EXHIBIT A

J. Mol. Biol. (1992) **227**, 776–798

The Repertoire of Human Germline V_H Sequences Reveals about Fifty Groups of V_H Segments with Different Hypervariable Loops

Ian M. Tomlinson[1,2], Gerald Walter[1,2], James D. Marks[1] Meirion B. Llewelyn[2] and Greg Winter[1,2]†

[1]*MRC Centre for Protein Engineering and*
[2]*MRC Laboratory of Molecular Biology*
Hills Road, Cambridge CB2 2QH, U.K.

(Received 6 February 1992; accepted 3 June 1992)

We have used the polymerase chain reaction and V_H family-based primers to clone and sequence 74 human germline V_H segments from a single individual and built a directory to include all known germline sequences. The directory contains 122 V_H segments with different nucleotide sequences, 83 of which have open reading frames. The directory indicates that the structural diversity of the germline repertoire for antigen binding is fixed by about 50 groups of V_H segments: each group encodes identical hypervariable loops. The directory should help in mapping the V_H locus, in estimating somatic mutation and V_H segment usage and in designing and constructing synthetic antibody libraries.

Keywords: human antibodies; heavy chain variable region; V_H; polymerase chain reaction

1. Introduction

Antibody architecture accommodates a wealth of structural diversity. Heavy and light chain variable domains (V_H and V_L) each consist of a β-sheet scaffold, surmounted by three antigen-binding loops (complementarity-determining regions, or CDRs‡; Kabut & Wu, 1971) of different lengths which are fleshed with a variety of different side-chains. The structural diversity of the loops can create binding sites of a variety of shapes, ranging from almost flat surfaces (Amit *et al.*, 1986) to deep cavities (Alzari *et al.*, 1990). Underpinning the structural diversity is a combinatorial genetic diversity. For V_H domains, it is generated by the assembly of V_H, D (diversity) and J_H (joining) segments. Two of the CDRs (1 and 2) are encoded by the V_H segment, and CDR3 by the 3′ end of the V_H segment, the D segment and the 5′ end of the J_H segment. With nucleotide addition (N-region diversity at the V_H-D and D-J_H joins), the use of different reading frames in the D segment, and the combination of different rearranged heavy and light chains, the diversity of primary antibody libraries is huge (for reviews, see Tonegawa, 1983; Winter & Milstein, 1991). During an immune response, the antibody variable regions are further diversified by somatic hypermutation, leading to higher affinity binding of the antigen (Berek & Milstein, 1988).

The human V_H, D and J_H segments have been mapped to band q32.33 of chromosome 14 (Croce *et al.*, 1979; Kirsch *et al.*, 1982), and recombine during B cell development. Each V_H segment encodes a 5′ hydrophobic leader peptide and between 95 and 101 amino acid residues of the mature domain flanked at the 3′ end by two recombination signals consisting of a highly conserved heptamer (5′-CACAGTG-3′), a 23-base-pair spacer and a less-conserved nonamer. The V_H segments have evolved by unequal crossing-over, conversion, duplication and deletion (Wysocki & Gefter, 1989; Walter *et al.*, 1990) and can be divided into six families on the basis of nucleotide homology of 80% or above (Kodaira *et al.*, 1986; Lee *et al.*, 1987; Shen *et al.*, 1987; Berman *et al.*, 1988; Humphries *et al.*, 1988; Buluwela & Rabbitts, 1988). The number of V_H segments per individual has most recently been estimated as about 76 (25 V_H1 segments, 5 V_H2 segments, 28 V_H3 segments, 14 V_H4 segments, 3 V_H5 segments and 1 V_H6 segment; Walter *et al.*, 1990), although these figures are likely to be an underestimate (Berman *et al.*, 1988; Walter *et al.*, 1990).

Earlier attempts to clone human V_H segments have involved constructing and probing large cosmid libraries, and have been aimed at mapping and sequencing the whole V_H locus, including pseudogenes (Kodaira *et al.*, 1986; Lee *et al.*, 1987; Berman *et al.*, 1988). In contrast, we set out to

† Author to whom all correspondence should be addressed.

‡ Abbreviations used: CDR, complementarity-determining region; PCR, polymerase chain reaction; FR, framework region; u.v., ultraviolet light.

0022–2836/92/190776–23 $08.00/0

Table 1

Family-specific primers for PCR amplification of the V_H exon

VH1 primers	
VH1 LEA EX1	5′-CCC AAG CTT CCA TGG ACT GGA CCT GGA G-3′
VH1 LEA EX2	5′-CCC AAG CTT TCA TGG GCT GGA CCT GCA A-3′
VH1 LEA IN	5′-CCC AAG CTT G(A,G)A (A,G)G(A,G) GAT T(G,T) (A,G,T) (G,T)TC CAG T-3′
VH1 LEA EX3	5′-CCC AAG CTT (T,C) (C,T) (C,T) (A,G)CA G(G,A) (T,C,A) (G,A) (C,T) (C,T,G) (C,T)A(C,T,G) (T,G)C-3′
VH1 FR1 (2–8)	5′-CCC AAG CTT (C,G,T)CA(G,A) (C,T)T(A,G,T) (G,T)T(G,A) (C,T)A(G,A) (T,C)C(T,G) G-3′
VH1 FR1 (17–22)	5′-CCC AAG CTT (T,A)C(A,G) G(T,C)G A(A,G) (G,A) (G,A)T(C,T) (T,A)CC TGC-3′
VH1 HEPT	5′-GGA ATT CT(C,G) TGG (G,T)TT (C,T)TC ACA CTG TG-3′
VH2 primers	
VH2 LEA	5′-CCC AAG CTT CTT CTC CAC AGG GGT CTT ATC-3′
VH2 HEPT	5′-GGA ATT CCA CTG TG(C,T) (C,G)CC GCG CAC A-3′
VH3 primers	
VH3 LEA1	5′-CCC AAG CTT T(A,T)(C,T) (A,G)TG TGG CA(A,G,C,T) TTT CTG A-3′
VH3 LEA2	5′-CCC AAG CTT T(A,T) (C,T) (A,G)T(C,G) TG(A,G) (A,C)A(A,G,C,T) TTT CTG A-3′
VH3 LEA3	5′-CCC AAG CTT GT(A,T) TGC A(A,G)G TG(C,T) CCA GTG T-3′
VH3 HEPT	5′-GGA ATT C(A,C)T G(A,G)C (C,T)TC CCC TC(A,G) CT(C,G) TG-3′
VH3 FR1	5′-CCC CCA AGC TTT GT(G,C) CAG (G,C)CT CTG G(A,G)T TC-3′
VH3 FR3	5′-GCT CTA GAG T(G,A)A (G,A)TC (T,G)GC C(T,C)T TCA C(A,G)G-3′
VH3 NON1	5′-GCT CTA GAG GTT TGT G(T,C)C (T,C)GG GC(G,T) CA-3′
VH4 primers	
VH4 LEA	5′-CCC AAG CTT CTG TTC ACA GGG GTC CTG TC-3′
VH4 HEPT	5′-GGA ATT CAC TCA CCT CCC CTC ACT GTG-3′
VH5 primers	
VH5 LEA	5′-CCC AAG CTT AGG TCA CAG AG(A,G) AGA A(C,T)G G-3′
VH5 HEPT	5′-GGA ATT CGC TGG TTT CTC TCA CTG TG-3′
VH6 primers	
VH6 LEA	5′-CCC AAG CTT TCA CAG CAG CAT TCA CAG A-3′
VH6 HEPT	5′-GGA ATT CCT GAC TTC CCC TCA CTG TG-3′

determine the repertoire of human V_H segments that contribute to the structural diversity of the V_H domain. We employed the polymerase chain reaction (PCR) (Saiki *et al.*, 1988) as a method of amplifying individual V_H segments. We designed family-specific primers for V_H segments based on the heptamer and part of the recombination spacer at the 3′ end of the V_H exon, and regions of the leader exon or intron at the 5′ end. Priming from the heptamer has been used to amplify mouse (Borghesi-Nicoletti & Schulze, 1991) and human (Sanz *et al.*, 1989*c*) V_H segments and has the advantage that since the heptamer is lost during recombination, rearranged V_H genes are not amplified.

2. Materials and Methods

(a) *Primer design*

Primers were designed (Table 1) for each of the 6 V_H families based on the sequences of published V_H segments (Kodaira *et al.*, 1986; Lee *et al.*, 1987; Berman *et al.*, 1988; Humphries *et al.*, 1988) and were located as shown in Fig. 1(a). Forward primers were based around the highly conserved heptamer recombination sequence, 5′-CACAGTG-3′. For 5 V_H families, published germline sequences were used, basing forward primers (VH1 HEPT, VH3 HEPT, VH4 HEPT, VH5 HEPT, VH6 HEPT) on the heptamer sequence and an additional 11 to 13 nucleotides from the recombination spacer. Degenerate nucleotides were incorporated to ensure the efficient priming of known germline genes from each V_H family, and *Eco*RI restriction sites were added for cloning. As germline V_H2 sequences were not available, the forward primer (VH2 HEPT) was designed using the sequence of the third framework (FR) region from a rearranged V_H2 gene, $V_{CE\text{-}1}$ (Takahashi *et al.*, 1984) adding 2 degenerate bases to substitute for those outside FR3, and then adding the conserved heptamer sequence. Family-specific back primers (VH1 LEA EX1, VH1 LEA EX2, VH1 LEA IN, VH1 LEA EX3, VH2 LEA, VH3 LEA1, VH3 LEA2, VH3 LEA3, VH4 LEA, VH5 LEA, VH6 LEA) were based on those parts of the leader exon and intron that are highly conserved within, but not between V_H families, again incorporating degeneracy where necessary (VH1 LEA EX1 and VH1 LEA EX2 were mixed in equal ratios and are referred to as VH1 LEA EX1/2). The back primers, VH1 FR1 (2–8) and VH1 FR1 (17–22), were subsequently designed using the sequences obtained with the first set of PCR primers. *Hin*dIII restriction sites were added to all back primers for cloning.

"Internal" primers for the V_H3 family were designed based on those regions of framework 1 (VH3 FR1) and CDR2-framework 3 (VH3 FR3) that display the greatest homology within the V_H3 family (see Fig. 2(b)). Since *Eco*RI restriction sites were noted in 2 published V_H3 pseudogenes ($V_{71\text{-}1}$ and $V_{71\text{-}3}$; Kodaira *et al.*, 1986) we changed the cloning site in the forward primer (VH3 FR3) to *Xba*I.

(b) *Preparation of genomic DNA*

Genomic DNA was isolated from peripheral white blood cells obtained from a healthy Caucasian donor, DP, using a method described by Perry & Carrell (1989). Briefly,

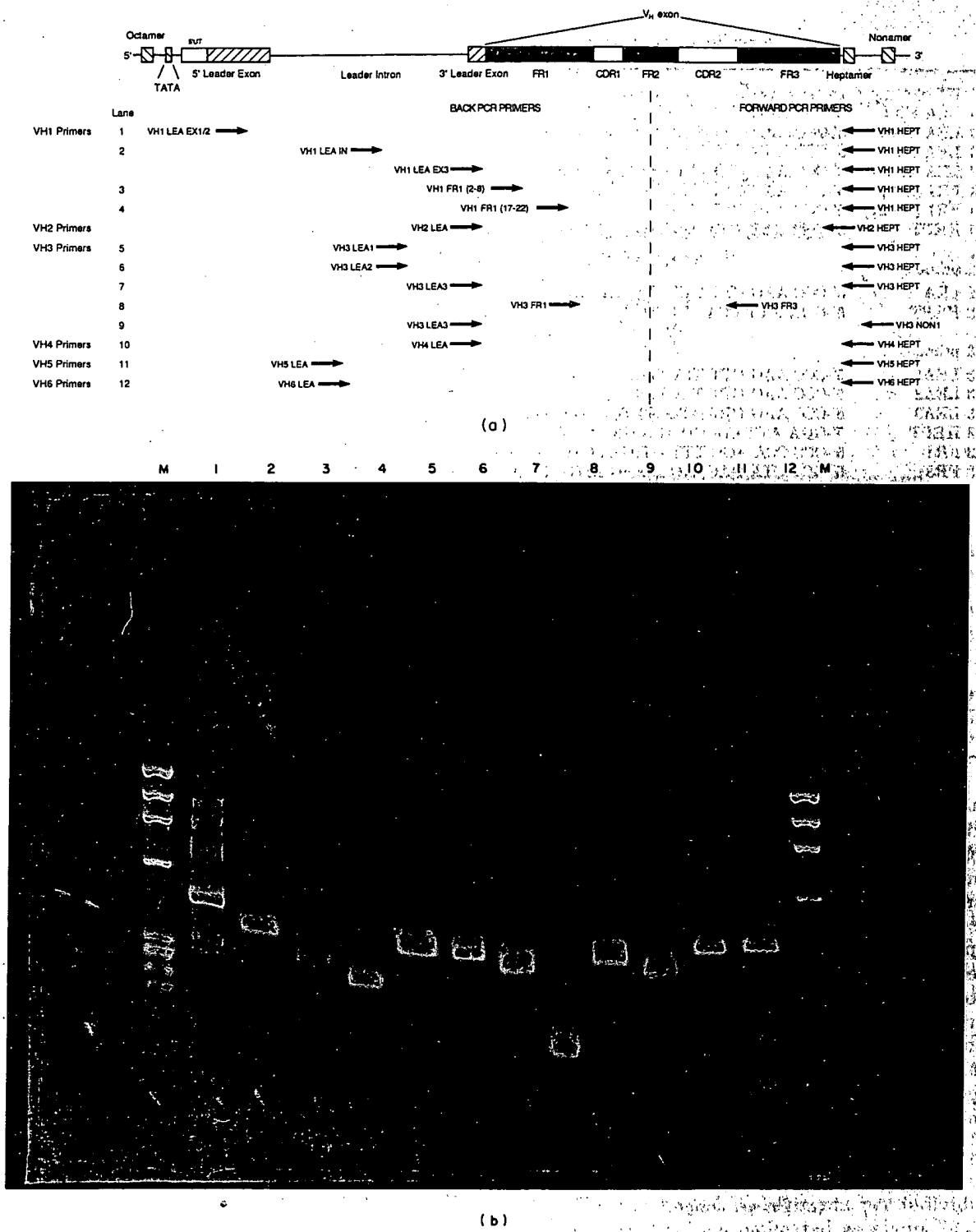


Figure 1. Family-specific primers for PCR amplification of the V_H exon. (a) Locations of the family-based PCR primers with respect to the V_H exon. FR, framework region; CDR, complementarity-determining region. Back primers were based in either the leader exon or intron or in framework 1 of the V_H segment. Forward primers were based around the heptamer and nonamer and at the junction of the CDR2 and framework 3. (b) PCR amplified genomic DNA from DP run on a 1·5% agarose gel. M, ϕX174 M_r markers; lanes 1 to 12, amplifications using the sets of primers depicted in (a).

9 ml whole blood was collected in 1 ml 3·8% (w/v) trisodium citrate (anticoagulant). The cells were lysed by adding the mixture to 90 ml ice-cold cell lysis buffer (0·32 M-sucrose, 1% Triton X-100, 5 mM-$MgCl_2$, 10 mM-Tris·HCl (pH 7·5)) and left on ice for 15 min. The nuclear pellet was isolated by centrifugation at 1000 *g* at 4 °C for 15 min and then resuspended in 4·5 ml Tris/EDTA (10 mM-Tris·HCl (pH 8·0), 1 mM-EDTA). The pellet was lysed using 10 ml nuclear lysis buffer (0·32 M-lithium acetate, 2% (w/v) SDS, 10 mM-Tris·HCl (pH 8·0), 1 mM-

EDTA), extracted twice with phenol/chloroform, once with chloroform and precipitated using ice-cold ethanol. Samples were resuspended in 500 μl of water and quantified by measuring their absorbance at 260 nm.

(c) *PCR amplification and sequencing*

Primers were synthesized on an Applied Biosystems (Foster City, CA) oligonucleotide synthesizer. Genomic DNA was amplified using the pairs of PCR primers (Fig. 1(a) and Table 1) in a Techne programmable Dri-Block PHC-1 thermal cycler (Cambridge, UK) with either Promega (Madison, WI) or Cetus (Perkin Elmer, Norwalk, CT) *Thermus aquaticus* (Taq) DNA polymerase. Reaction mixtures (50 μl) were prepared containing 25 pmol of each primer, 5 to 10 μg of genomic DNA, 2·5 units of Taq polymerase, 200 μM (each) dNTPs and the recommended buffer (Promega: 50 mM-KCl, 10 mM-Tris·HCl (pH 8·8), 1·5 mM-$MgCl_2$, 0·1% Triton X-100; Cetus: 50 mM-KCl, 10 mM-Tris·HCl (pH 8·3), 1·5 mM-$MgCl_2$, 0·001% (w/v) gelatin). The reaction mixture was overlaid with paraffin oil and 30 cycles of amplification were performed. Each cycle consisted of denaturation (94 °C for 1 min), annealing (55 °C for 1 min) and extension (72 °C for 2 min). At the end of 30 cycles, there was a final extension at 65 °C for 5 min. The product was analysed by running 5 μl on a 1·5% (w/v) agarose gel. The remainder was extracted with phenol/chloroform, precipitated with ethanol and digested with restriction enzymes *Hin*dIII and *Eco*RI (or *Xba*I). A band of the expected size was cut from a 1·5% low melting point agarose gel and then purified by adsorption onto glassmilk using Geneclean II (Bio 101, La Jolla, CA) or by electroelution followed by precipitation with ethanol.

The product was ligated into M13-K19 (Carter *et al.*, 1985) that had been digested with *Hin*dIII and *Eco*RI (or *Xba*I). The ligation mix was used to transform *E. coli* BMH 71-18 cells (Gronenborn, 1976) by electroporation (Dower *et al.*, 1988) using the Bio-Rad (Richmond, CA) Gene Pulser and plated on TYE plates (Miller, 1972). Single-stranded template from selected plaques was prepared and sequenced using the dideoxy chain termination method (Sanger *et al.*, 1977) and modified T7 DNA polymerase (Sequenase II, United States Biochemical Corp., Cleveland, Ohio). The sequence was read in one direction and compressions resolved using deoxyinosine triphosphate (Mills & Kramer, 1979).

Several precautions were taken to avoid cross-contamination. PCR reaction mixes were subjected to high intensity short-wave u.v. radiation (Amplirad, Genetic Research Instrumentation, Dunmow, Essex, U.K.) for 5 min before adding genomic DNA to destroy any DNA contamination. Negative controls (no genomic DNA added) were always included in all amplifications to check for DNA contamination. Independent amplifications with identical sets of primers were undertaken simultaneously to avoid clones isolated from one amplification contaminating the next. In all cases we imposed the requirement that each germline V_H segment was seen in at least 2 independent amplifications.

(d) *Probing*

Oligonucleotide probes, 17 to 21 nucleotides in length (Table 2) were designed as described in Results, and synthesized as above. Phage plaques were picked onto duplicate TYE plates and grown as colonies for 30 h at 37 °C. (Plaques that should hybridize to the probes were always included as positive controls.) The colonies were lifted onto Hybond nylon filters (Amersham Int., Amersham, U.K.), denatured in 5% (w/v) SDS, 2 × SSC (300 mM-NaCl, 30 mM-trisodium citrate, pH 7·0) for 2 min, baked in a microwave oven for 2·5 min and autocrosslinked by short-wave u.v. (Stratalinker: Stratagene, La Jolla, CA) (Buluwela *et al.*, 1989). Filters were prehybridized for 20 min at 42 °C in 15 ml hybridization solution (1 M-NaCl, 1 × Denhardt's (0·02% Ficoll, 0·02% polyvinylpyrrolidone, 0·02% bovine serum albumin), 100 mM-Tris·HCl (pH 7·5), 6·25 mM-EDTA, 1 mM-sodium pyrophosphate, 0·5% Nonidet P40, 0·006% rATP, 0·02% brewers' yeast tRNA) using a Techne HB-1 Hybridiser (Cambridge, U.K.).

For probing, 15 pmol of oligonucleotide were phosphorylated with 30 μCi [^{32}P]dATP for 30 min using 2 units of polynucleotide kinase (New England Biolabs, Beverly, MA) in 30 μl 50 mM-Tris·HCl (pH 7·5), 10 mM-$MgCl_2$, 1 mM-dithiothreitol, and incorporation of ^{32}P checked by electrophoresis of the oligonucleotide on an 18% (w/v) polyacrylamide gel. The probe was added to the hybridization solution, and the filters were hybridized at 42 °C for 2 h and then washed with 40 ml 6 × SSC (see above), 0·1% SDS, 0·1% sodium pyrophosphate at this temperature for 15 min and then for 20 min with 40 ml 3 M-TMACl (tetramethylammonium chloride) in 50 mM-Tris·HCl (pH 8·0), 0·1% SDS and 2 mM-EDTA (Wood *et al.*, 1985) at 59 °C (17-mer), 61 °C (18-mer), 63 °C (19-mer) or 67 °C (21-mer). Filters were dried and exposed to Kodak Fast Film overnight using an intensifying screen at −70 °C. Filters were recycled by washing at 80 to 90 °C for 5 min in 2 × SSC and could be probed several times without loss of signal.

(e) *Compilation of germline and rearranged V_H database*

DNA sequences were aligned and translated by a sequence analysis program (MacVector, IBI Kodak, New Haven, CT). In order to compile a comprehensive database of both human germline and rearranged V_H sequences we searched MedLine (U.S. National Library of Medicine), GenBank (IntelliGenetics Inc., Real Mountain View, CA) and Kabat (Proteins of Immunological Interest, Kabat *et al.*, 1991) databases (for references, see Figs 2 and 3) and incorporated our own data. Rearranged genes were assigned to their closest germline counterparts by the presence of specific motifs in the protein sequence indicative of a particular V_H segment or by maximum homology of the nucleotide sequences (using MacVector).

3. Results

(a) *Strategy*

We designed family-specific PCR primers based on sequences from the literature and amplified, cloned and sequenced germline V_H segments from our donor DP. Nucleotide sequences were aligned and taken as confirmed when seen as identical in two independent amplifications. Genes which remained unconfirmed in phase 1 were probed for with ^{32}P-labelled oligonucleotides and sequenced in phase 2.

(b) *Phase 1: PCR amplification and sequencing of random clones*

Genomic DNA was amplified using sets of family-based primers. The majority of primer combinations

Table 2

Oligonucleotide probes used for identification of germline V_H segments

V_H1 family	
DP-1	5′-AGT AAT ACG TGG CCG TG-3′
DP-1/8	5′-TGT GCC ACC ACT GTT AG-3′
DP-2	5′-CCA CTG CCA ACG ACG AT-3′
DP-3/5/24	5′-ATT GTT TCA CCA TCT TC-3′
DP-4	5′-TGC AGG TAG CGG TAG GT-3′
DP-4	5′-ACC ATT GAA AGG TGT GA-3′
DP-5	5′-TGG ATA ATT CAG TGA GG-3′
DP-6	5′-ACT GTG TAA AGT ATT TG-3′
DP-7/22	5′-CAG TGC ATA TAG TAG CT-3′
DP-8	5′-TCG TCA GAT CTC AGC CT-3′
DP-9/10	5′-CCA GCT GAT AGC ATA GC-3′
DP-9/21	5′-GGT TCC CAG TGT TGG TG-3′
DP-10	5′-TGC TGT ACC AAA GAT AG-3′
DP-11	5′-AGG TGT ATC CAC AAG TCT-3′
DP-12	5′-ATC ACT AGG GCA CAC CAA-3′
DP-13	5′-ACA TTG GGT TCA CCA GGG-3′
DP-14/22	5′-TGT GTT ACC ATT GTA AG-3′
DP-15	5′-AGT TGA TAT CAT AAC TG-3′
DP-16/17/20	5′-TTG CCA GAG TAG CTC CC-3′
DP-18	5′-GAT CTG AAG ACA CGC CG-3′
DP-19	5′-GAC TAC ACC AGT TGG AC-3′
DP-19	5′-GTT CAT AAA GTA GTC GG-3′
DP-19	5′-TGC TCG AAG ATG TGT CC-3′
DP-19/23/25	5′-GTG TTA CCA TTG CCA GC-3′
DP-21	5′-CAA CTC AGA CCC AGA TT-3′
DP-22	5′-CGG CCA TGT CGT CAG AT-3′
DP-23	5′-GCA TAA AGT TGT TGG TG-3′
DP-24	5′-CCC AGG TTT CCT CAC CT-3′
DP-1/7/8/10/14/19/21/22/23/25	5′-CAC TGT GTC TCT CGC AC-3′
Rearranged gene probes	
333, 1H1, etc.	5′-CAG TGT ATG GTG GAG TCA C-3′
VDJ191	5′-AGT CAG GGC ATG ATT ATT A-3′
39-1	5′-GCC CAC ACC CAC TCC ACT AGT-3′
41-1	5′-GCC CAC ACC CCC TCC ACT AGT-3′
V_H3 family	
DP-29	5′-TTG TTT CTA GTA CGG CCA A-3′
DP-30	5′-TTC TTA TTA AAC CTA CCA A-3′
DP-31	5′-CAC TAT TCC AAC TAA TAC C-3′
DP-32	5′-GTG CTA CCA CCA TTC CAA T-3′
DP-33	5′-CAC CAT CCC AAC TAA TAA G-3′
DP-36	5′-AGC TTT GCT TTT AAT ACA G-3′
DP-37	5′-AGC TTT GCT TTT AAT ACG G-3′
DP-41	5′-GTG CAT GCC ATA GTT ACT G-3′
DP-42	5′-TAC CAC CGC TAT AAA TAA C-3′
DP-44/45	5′-GTG CCA CCA CCA GTA CCA A-3′
DP-44/45/46/61	5′-CAG TGC ATA GCA TAG CTA C-3′
DP-47	5′-CCA CTA CCA CTA ATA GCT G-3′
DP-49/50	5′-CAG TGC ATG CCA TAG CTA C-3′
DP-46/49	5′-TCA TAT GAT ATA ACT GCC A-3′
DP-50	5′-TCA TAC CAT ATA ACT GCC A-3′
DP-51	5′-CAG TTC ATG CTA TAG CTA C-3′
DP-52	5′-CAG TGC AGA ACA TAG CTA C-3′
DP-53	5′-CCA TCA CTA TTA ATA CGT G-3′
DP-54	5′-TTC CAT CTT GCT TTA TGT T-3′
DP-55/56	5′-CCC CAT TAG GAT TAA CTT G-3′
DP-58	5′-AGT TCA TTT CAT AAC TA-3′
DP-59	5′-CAG TTC ATG TCA CTG TTA C-3′
DP-60	5′-GTA GCC ATA GCA CGC ACT G-3′
DP-61	5′-ACC CCC ATT ACT ACT AAT A-3′
V_H4 family	
$V_{2\text{-}1}$	5′-GCC CCA GTA GTA ACT ACT ACT-3′
V_{58}	5′-GTA GTA ACC ACT GAC GGA C-3′
V_{11}	5′-AGT TGG GGT TCC CAC TAT G-3′
V_{79}	5′-GGT CCC CGG AGG CTT CAC C-3′

produced good intensity PCR bands, as is shown in Figure 1(b), but amplifications using VH1 EX3/VH1 HEPT and VH2 LEA/VH2 HEPT were variable and hence are not shown. Initially, 596 random clones were sequenced (V_H1 family (170), V_H2 family (120), V_H3 family (150), V_H4 family (120), V_H5 family (24) and V_H6 family (12)). With one exception (one V_H5 gene found in a V_H1 library), the primers proved family-specific. This initial round of sequencing established 35 V_H sequences (including pseudogenes) that were identical in at least *two independent* PCR amplifications (V_H1 family (12), V_H2 family (3), V_H3 family (8), V_H4 family (10), V_H5 family (1) and V_H6 family (1)) and by this criterion correspond to germline V_H segments.

Many sequences were unconfirmed due to single nucleotide differences between clones from independent amplifications, presumably due to errors introduced by the Taq polymerase. The 61 single base changes seen per 100 sequences for the V_H1 and V_H3 families correspond to 7×10^{-5} changes/nucleotide per cycle, which is consistent with the Taq polymerase error rate suggested by Maruyama (1990).

Other sequences, never confirmed in independent amplifications (but sometimes found in more than one clone from the same amplification), consisted of two parts, each of which could be aligned to different V_H segments. As became clear on probing (see below), these sequences arose from partially extended fragments reannealing to a different segment after heat-denaturation. This phenomenon, termed "PCR cross-over", has also been seen in the detection of homologous recombinants (Frohman & Martin, 1990) and in the amplification of prepro-insulin cDNA (Shuldiner *et al.*, 1989) and in this study accounted for 10% of all V_H1 and V_H3 clones sequenced.

For the smaller V_H families (V_H2, V_H4, V_H5, V_H6), all sequences were confirmed in phase 1, or could be explained by PCR artifacts. But many sequences from the V_H1 and V_H3 families remained unconfirmed, requiring systematic probing of a larger number of clones.

(c) *Phase 2: probing and directed sequencing*

With the V_H1 primers, 42 different sequences (excluding obvious PCR errors caused by single base substitutions) were obtained in phase 1. Only 12 of these sequences were identical in at least two independent amplifications. Therefore, motif-specific probes were designed (Table 2) such that each probe would identify a group of different V_H1 clones with a particular sequence motif. Hence, when each clone

was probed in turn with each of the 29 probes, it could be distinguished by its "fingerprint", i.e. the set of sequence motifs that it contains. Thus, 1750 clones from independent amplifications using the five V_H1-based primer combinations (Fig. 1) were regridded and hybridized with the 29 probes. Clones that appeared to confirm a sequence from phase 1 by "fingerprinting" were sequenced. In this way a further 11 V_H1 sequences were confirmed and only two new (pseudo)genes (DP-17, DP-20) were discovered. Nineteen of the original 42 sequences could not be confirmed by probing, but 18 of these could be attributed to "PCR cross-over".

For the majority of unconfirmed sequences in the V_H3 family, we designed gene-specific probes (17- and 19-mers, Table 2), except in the case of DP-46/DP-49, where three probes were necessary for identification, and DP-44/45 and DP-56/57, where discrimination between the two in each pair was not possible. Probes were centred on the region of greatest heterogeneity within a CDR and therefore a single probe (with the above exceptions) could identify a single V_H segment. Thus, 1100 clones taken from independent amplifications with the three sets of V_H3 leader/heptamer-based primers (Fig. 1) were hybridized in turn with the 21 probes and a further 22 V_H3 segments were confirmed by directed sequencing. The remaining unconfirmed sequences could be attributed to PCR artifacts.

We also designed "internal" VH3 primers (VH3 FR1 and VH3 FR3) based on sequence data from phase 1 and phase 2. Genomic DNA from DP was amplified as before, and 48 randomly selected clones were sequenced and confirmed, when necessary, in two independent amplifications by probing and directed sequencing. Only seven new V_H segments were obtained, three of which appeared to be fragmented pseudogenes with less than 60% homology to any known V_H segment. Two sequences had been published before and have unusual heptamer sequences (DP-59/V_H19 and DP-62/$V_{71\text{-}1}$, respectively), and the other two sequences were new (DP-60 and DP-61).

To isolate full length versions of genes DP-59 to DP-61, which have open reading frames, we designed a primer (VH3 NON1) based on nonamer sequences of V_H segments reviewed by Pascual & Capra (1991). Amplifications of genomic DNA were performed using VH3 LEA3 and VH3 NON1, and the resulting fragments were cloned, regridded and probed with oligonucleotides specific for DP-59, DP-60 and DP-61. DP-59 and DP-60 were isolated from independent PCR amplifications, and shown to have unusual heptamer sequences. A full length version of DP-61 was not found in this library.

We also attempted to confirm additional germline V_H segments reported in the literature and germline analogues of published rearranged genes. Using the V_H family-specific primers (Table 1) to amplify and clone germline V_H2, V_H3 and V_H4 segments, we probed (Table 2) for the germline V_H segments V_{11}, V_{58}, V_{79} and $V_{2\text{-}1}$ (Lee *et al.*, 1987), rearranged V_H genes 39-1, 41-1 (Deane & Norton, 1990), VDJ191 (Mensink *et al.*, 1986) and 333, 1H1, 2C12, 2A12, 1B11, 112, 115 and 126 (Cleary *et al.*, 1986) (rearranged genes were probed for at low stringency, i.e. TMACl wash at 50°C). None of these genes was identified in our libraries.

(d) *Sequence directory*

The 74 germline V_H segments (25 V_H1 segments, 3 V_H2 segments, 34 V_H3 segments, 10 V_H4 segments, 1 V_H5 segment and 1 V_H6 segment) cloned and sequenced by us are prefixed "DP", the initials of our donor and are denoted by running numbers. Of these, 51 have open reading frames and 23 contain either frame shifts or stop codons and are therefore considered to be pseudogenes. We have also included sequences of germline V_H segments published by others. The protein and nucleotide sequences of all 83 germline V_H segments with open reading frames are given in Figure 2(a) and (b), respectively, and nucleotide sequences of the 39 germline V_H segments with interrupted reading frames (either frame shifts or stop codons) in Figure 2(c). In Figure 2(b), the nucleotide sequences in each family have been aligned to a sequence with an open reading frame, 21-2 (V_H1 family), $V_{II\text{-}5}$ (V_H2 family), 12-2 (V_H3 family), $V_{71\text{-}2}$ (V_H4 family), VH251 (V_H5 family), V_H-VI (V_H6 family). The same sequences were used to align the pseudogenes in Figure 2(c).

"fl-pl" is a V_H segment described by Olee *et al.* (1991), which was seen in amplifications of genomic DNA from two individuals, Fer and Pla. The V_H segments hv3005b54, hv3019b13, hv3019b18 (Olee *et al.*, 1991), $V_H4.12$, $V_H4.14$, $V_H4.15$ (Sanz *et al.*, 1989*c*) are genes amplified by PCR, but not confirmed either by probing, independent amplifications, a rearranged sequence or by independent work. These sequences may be the result of PCR artifacts and have therefore been excluded from Figure 2.

Within each family, protein sequences are arranged alphabetically by the amino acid residues (single letter code) of CDR1 and where these are identical by CDR2 (Fig. 2(a)). Sequences with minor framework differences, which could include allelic differences, are therefore adjacent. Sequences with identical encoded CDRs 1 and 2 are grouped with brackets (these also have identical H1 and H2 hypervariable loops, as defined by Chothia *et al.* (1992), except in the case of 21-2/3-1/DP-7 and HG3; and $V_H4.11$/DP-71, $V_{71\text{-}4}$ and $V_H4.16$). The canonical structure classes of H1 (CDR1) and H2 (CDR2) (Chothia & Lesk, 1987; Chothia *et al.*, 1989, 1992) are shown, and those sequences that may be defective on structural grounds are marked with an *X* (see Chothia *et al.*, 1992). The canonical structure class of DP-61 is unknown.

V_H segments that have heptamers other than the conserved 5′-CACAGTG-3′ motif are marked *H*. The nonamer is generally conserved within each

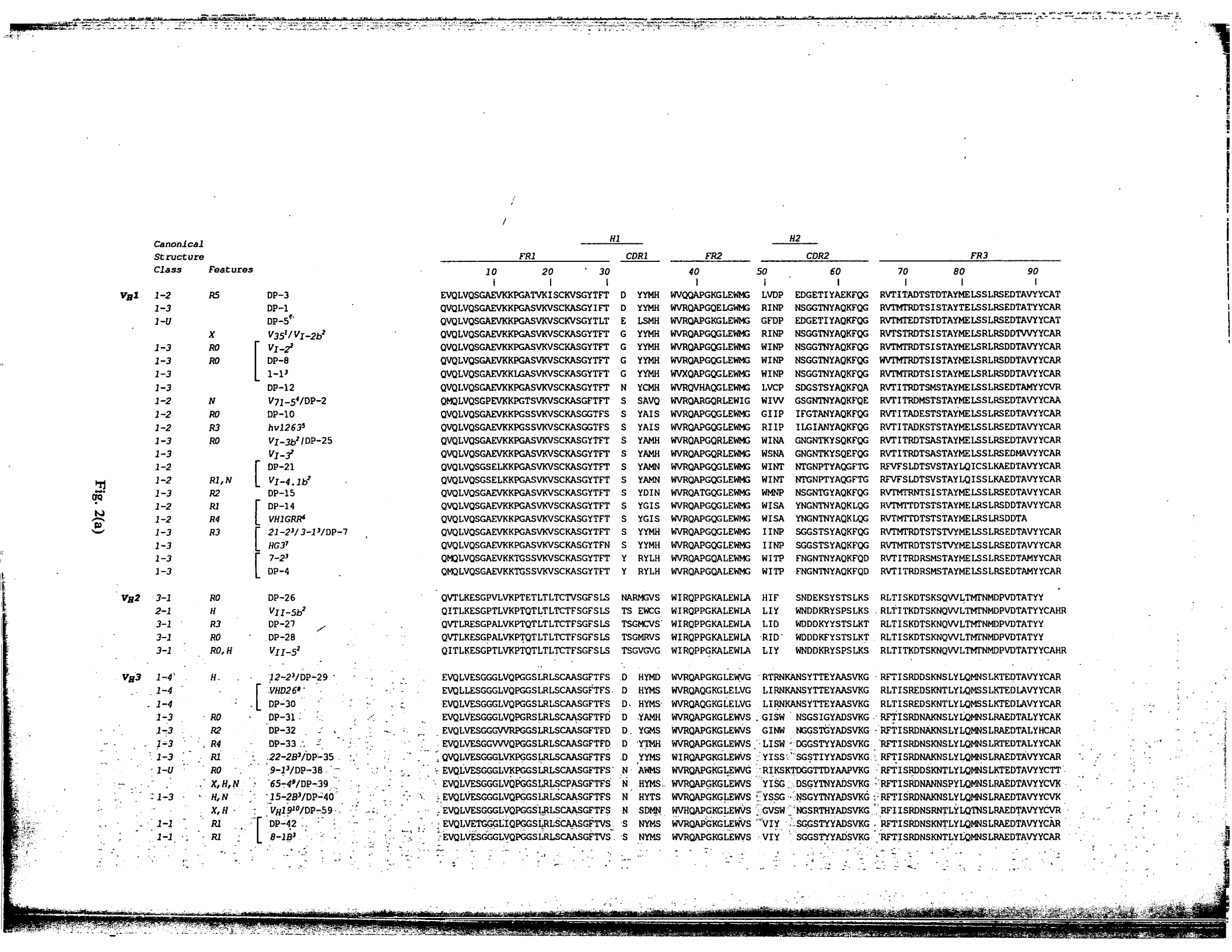

H1
H2
Canonical Structure Class
Features
FR1
CDR1
FR2
CDR2
FR3
10 20 30 40 50 60 70 80 90
VH1 1-2 R5 DP-3 EVQLVQSGAEVKKPGATVKISCKVSGYTFT D YYMH WVQQAPGKGLEWMG LVDP EDGETIYAEKFQG RVTITADTSTDTAYMELSSLRSEDTAVYYCAT
1-3 DP-1 QVQLVQSGAEVKKPGASVKVSCKASGYIFT D YYMH WVRQAPGQELGWMG RINP NSGGTNYAQKFQG RVTMTRDTSISTAYTELSSLRSEDTATYYCAR
1-U DP-5[6] QVQLVQSGAEVKKPGASVKVSCKVSGYTLT E LSMH WVRQAPGKGLEWMG GFDP EDGETIYAQKFQG RVTMTEDTSTDTAYMELSSLRSEDTAVYYCAT
X V35[1]/VI-2b[2] QVQLVQSGAEVKKPGASVKVSCKASGYTFT G YYMH WVRQAPGQGLEWMG RINP NSGGTNYAQKFQG RVTSTRDTSISTAYMELSRLRSDDTVVYYCAR
1-3 R0 VI-2[2] QVQLVQSGAEVKKPGASVKVSCKASGYTFT G YYMH WVRQAPGQGLEWMG WINP NSGGTNYAQKFQG RVTMTRDTSISTAYMELSRLRSDDTAVYYCAR
1-3 R0 DP-8 QVQLVQSGAEVKKPGASVKVSCKASGYTFT G YYMH WVRQAPGQGLEWMG WINP NSGGTNYAQKFQG WVTMTRDTSISTAYMELSRLRSDDTAVYYCAR
1-3 1-1[3] QVQLVQSGAEVKKLGASVKVSCKASGYTFT G YYMH WVXQAPGQGLEWMG WINP NSGGTNYAQKFQG RVTMTRDTSISTAYMELSRLRSDDTAVYYCAR
1-3 DP-12 QVQLVQSGAEVKKPGASVKVSCKASGYTFT N YCMH WVRQVHAQGLEWMG LVCP SDGSTSYAQKFQA RVTITRDTSMSTAYMELSSLRSEDTAMYYCVR
1-2 N V71-5[4]/DP-2 QMQLVQSGPEVKKPGTSVKVSCKASGFTFT S SAVQ WVRQARGQRLEWIG WIVV GSGNTNYAQKFQE RVTITRDMSTSTAYMELSSLRSEDTAVYYCAA
1-2 R0 DP-10 QVQLVQSGAEVKKPGSSVKVSCKASGGTFS S YAIS WVRQAPGQGLEWMG GIIP IFGTANYAQKFQG RVTITADESTSTAYMELSSLRSEDTAVYYCAR
1-2 R3 hv1263[5] QVQLVQSGAEVKKPGSSVKVSCKASGGTFS S YAIS WVRQAPGQGLEWMG RIIP ILGIANYAQKFQG RVTITADKSTSTAYMELSSLRSEDTAVYYCAR
1-3 R0 VI-3b[2]/DP-25 QVQLVQSGAEVKKPGASVKVSCKASGYTFT S YAMH WVRQAPGQRLEWMG WINA GNGNTKYSQKFQG RVTITRDTSASTAYMELSSLRSEDTAVYYCAR
1-3 VI-3[2] QVQLVQSGAEVKKPGASVKVSCKASGYTFT S YAMH WVRQAPGQRLEWMG WSNA GNGNTKYSQEFQG RVTITRDTSASTAYMELSSLRSEDMAVYYCAR
1-2 DP-21 QVQLVQSGSELKKPGASVKVSCKASGYTFT S YAMN WVRQAPGQGLEWMG WINT NTGNPTYAQGFTG RFVFSLDTSVSTAYLQICSLKAEDTAVYYCAR
1-2 R1,N VI-4.1b[2] QVQLVQSGSELKKPGASVKVSCKASGYTFT S YAMN WVRQAPGQGLEWMG WINT NTGNPTYAQGFTG RFVFSLDTSVSTAYLQISSLKAEDTAVYYCAR
1-3 R2 DP-15 QVQLVQSGAEVKKPGASVKVSCKASGYTFT S YDIN WVRQATGQGLEWMG WMNP NSGNTGYAQKFQG RVTMTRNTSISTAYMELSSLRSEDTAVYYCAR
1-2 R1 DP-14 QVQLVQSGAEVKKPGASVKVSCKASGYTFT S YGIS WVRQAPGQGLEWMG WISA YNGNTNYAQKLQG RVTMTTDTSTSTAYMELRSLRSDDTAVYYCAR
1-2 R4 VH1GRR[6] QVQLVQSGAEVKKPGASVKVSCKASGYTFT S YGIS WVRQAPGQGLEWMG WISA YNGNTNYAQKLQG RVTMTTDTSTSTAYMELRSLRSDDTA
1-3 R3 21-2[3]/3-1[3]/DP-7 QVQLVQSGAEVKKPGASVKVSCKASGYTFT S YYMH WVRQAPGQGLEWMG IINP SGGSTSYAQKFQG RVTMTRDTSTSTVYMELSSLRSEDTAVYYCAR
1-3 HG3[7] QVQLVQSGAEVKKPGASVKVSCKASGYTFN S YYMH WVRQAPGQGLEWMG IINP SGGSTSYAQKFQG RVTMTRDTSTSTVYMELSSLRSEDTAVYYCAR
1-3 7-2[3] QMQLVQSGAEVKKTGSSVKVSCKASGYTFT Y RYLH WVRQAPGQALEWMG WITP FNGNTNYAQKFQD RVTITRDRSMSTAYMELSSLRSEDTAMYYCAR
1-3 DP-4 QMQLVQSGAEVKKTGSSVKVSCKASGYTFT Y RYLH WVRQAPGQALEWMG WITP FNGNTNYAQKFQD RVTITRDRSMSTAYMELSSLRSEDTAMYYCAR
VH2 3-1 R0 DP-26 QVTLKESGPVLVKPTETLTLTCTVSGFSLS NARMGVS WIRQPPGKALEWLA HIF SNDEKSYSTSLKS RLTISKDTSKSQVVLTMTNMDPVDTATYY
2-1 H VII-5b[2] QITLKESGPTLVKPTQTLTLTCTFSGFSLS TS EWCG WIRQPPGKALEWLA LIY WNDDKRYSPSLKS RLTITKDTSKNQVVLTMTNMDPVDTATYYCAHR
3-1 R3 DP-27 QVTLRESGPALVKPTQTLTLTCTFSGFSLS TSGMCVS WIRQPPGKALEWLA LID WDDDKYYSTSLKT RLTISKDTSKNQVVLTMTNMDPVDTATYY
3-1 R0 DP-28 QVTLKESGPALVKPTQTLTLTCTFSGFSLS TSGMRVS WIRQPPGKALEWLA RID WDDDKFYSTSLKT RLTISKDTSKNQVVLTMTNMDPVDTATYY
3-1 R0,H VII-5[2] QITLKESGPTLVKPTQTLTLTCTFSGFSLS TSGVGVG WIRQPPGKALEWLA LIY WNDDKRYSPSLKS RLTITKDTSKNQVVLTMTNMDPVDTATYYCAHR
VH3 1-4 H 12-2[3]/DP-29 EVQLVESGGGLVQPGGSLRLSCAASGFTFS D HYMD WVRQAPGKGLEWVG RTRNKANSYTTEYAASVKG RFTISRDDSKNSLYLQMNSLKTEDTAVYYCAR
1-4 VHD26[8] EVQLLESGGGLVQPGGSLRLSCAASGFTFS D HYMS WVRQAQGKGLELVG LIRNKANSYTTEYAASVKG RLTISREDSKNTLYLQMSSLKTEDLAVYYCAR
1-4 DP-30 EVQLVESGGGLVQPGGSLRLSCAASGFTFS D HYMS WVRQAQGKGLELVG LIRNKANSYTTEYAASVKG RLTISREDSKNTLYLQMSSLKTEDLAVYYCAR
1-3 R0 DP-31 EVQLVESGGGLVQPGRSLRLSCAASGFTFD D YAMH WVRQAPGKGLEWVS GISW NSGSIGYADSVKG RFTISRDNAKNSLYLQMNSLRAEDTALYYCAK
1-3 R2 DP-32 EVQLVESGGGVVRPGGSLRLSCAASGFTFD D YGMS WVRQAPGKGLEWVS GINW NGGSTGYADSVKG RFTISRDNAKNSLYLQMNSLRAEDTALYHCAR
1-3 R4 DP-33 EVQLVESGGVVQPGGSLRLSCAASGFTFD D YTMH WVRQAPGKGLEWVS LISW DGGSTYYADSVKG RFTISRDNSKNSLYLQMNSLRTEDTALYYCAK
1-3 R1 22-2B[3]/DP-35 QVQLVESGGGLVKPGGSLRLSCAASGFTFS D YYMS WIRQAPGKGLEWVS YISS SGSTIYYADSVKG RFTISRDNAKNSLYLQMNSLRAEDTAVYYCAR
1-U R0 9-1[3]/DP-38 EVQLVESGGGLVKPGGSLRLSCAASGFTFS N AWMS WVRQAPGKGLEWVG RIKSKTDGGTTDYAAPVKG RFTISRDDSKNTLYLQMNSLKTEDTAVYYCTT
X,H,N 65-4[9]/DP-39 EVQLVESGGGLVQPGGSLRLSCPASGFTFS N HYMS WVRQAPGKGLEWVS YISG DSGYTNYADSVKG RFTISRDNANNSPYLQMNSLRAEDTAVYYCVK
1-3 H,N 15-2B[3]/DP-40 EVQLVESGGGLVQPGGSLRLSCAASGFTFS N HYTS WVRQAPGKGLEWVS YSSG NSGYTNYADSVKG RFTISRDNAKNSLYLQMNSLRAEDTAVYYCVK
X,H VH19[10]/DP-59 EVQLVESGGGLVQPGGSLRLSCAASGFTFS N SDMN WVHQAPGKGLEWVS GVSW NGSRTHYADSVKG RFIISRDNSRNTLYLQTNSLRAEDTAVYYCVR
1-1 R1 DP-42 EVQLVETGGGLIQPGGSLRLSCAASGFTVS S NYMS WVRQAPGKGLEWVS VIY SGGSTYYADSVKG RFTISRDNSKNTLYLQMNSLRAEDTAVYYCAR
1-1 R1 8-1B[3] EVQLVESGGGLVQPGGSLRLSCAASGFTVS S NYMS WVRQAPGKGLEWVS VIY SGGSTYYADSVKG RFTISRDNSKNTLYLQMNSLRAEDTAVYYCAR

Fig. 2(a)

		X	65-2[9]/DP-44	EVQLVHSGGGLVQPGGSLRLSCAGSGFTFS	S	YAMH	WVRQAPGKGLEWVS	AIG	TGGGTYYADSVKG	RFTISRDNAKNSLYLQMNSLRAEDMAVYYCAR
	1-1		DP-45	EVQLVQSGGGLVQPGGSLRLSCAGSGFTFS	S	YAMH	WVRQAPGKGLEWVS	AIG	TGGGTYYADSVKG	RFTISRDNAKNSLYLQMNSLRAEDMAVYYCAR
	1-3	R10	f1-p1[11]	EVQLVESGGGLVQPGGSLRLSCSASGFTFS	S	YAMH	WVRQAPGKGLEYVS	AISS	NGGSTYYADSVKG	RFTISRDNSKNTLYVQMSSLRAEDTAVYYCVR
			DP-61	TFS	S	YAMH	WVRQAPGKGLEYVS	AISS	NGGSTYYAD	
	1-3		hv3005[12]	QVQLVESGGGVVQPGRSLRLSCAASGFTFS	S	YAMH	WVRQAPGKGLEWVA	VISY	DGSNKYYADSVKG	RFTISRDNSKNTLYLQMNSLRAEDTAVYYCAR
	1-3	R0	hv3005f3[11]	QVQLVESGGGVVQPGRSLRLSCAASGFTFS	S	YAMH	WVRQAPGKGLEWVA	VISY	DGSNKYYADSVKG	RFTISRDNSKNTLYLQMNSLRAEDTAVYYCAR
	1-3	R0	GL-SJ2[13]/DP-46	QVQLVESGGGVVQPGRSLRLSCAASGFTFS	S	YAMH	WVRQAPGKGLEWVA	VISY	DGSNKYYADSVKG	RFTISRDNSKNTLYLQMNSLRAEDTAVYYCAR
	1-3	R0	VH26[14]/DP-47	EVQLLESGGGLVQPGGSLRLSCAASGFTFS	S	YAMS	WVRQAPGKGLEWVS	AISG	SGGSTYYADSVKG	RFTISRDNSKNTLYLQMNSLRAEDTAVYYCAR
	1-1	R0	13-2[3]/DP-48	EVQLVESGGGLVQPGGSLRLSCAASGFTFS	S	YDMH	WVRQATGKGLEWVS	AIG	TAGDTYYPGSVKG	RFTISRDNSKNTLYLQMNSLRAEDTAVYYCAK
	1-3	R0	DP-58	EVQLVESGGGLVQPGGSLRLSCAASGFTFS	S	YEMN	WVRQAPGKGLEWVS	YISS	SGSTIYYADSVKG	RFTISRDNAKNSLYLQMNSLRAGDTAVYYCAR
	1-3	R0	1.9III[3]/DP-49	QVQLVESGGGVVQPGRSLRLSCAASGFTFS	S	YGMH	WVRQAPGKGLEWVA	VISY	DGSNKYYADSVKG	RFTISRDNAKNSLYLQMNSLRAEDTAVYYCAR
	1-3	R0	3019b9[11]/DP-50	QVQLVESGGGVVQPGRSLRLSCAASGFTFS	S	YGMH	WVRQAPGKGLEWVA	VIWY	DGSNKYYADSVKG	RFTISRDNSKNTLYLQMNSLRAEDTAVYYCAK
	1-3	R3	DP-51	EVQLVESGGGLVQPGGSLRLSCAASGFTFS	S	YSMN	WVRQAPGKGLEWVS	YISS	SSSTIYYADSVKG	RFTISRDNAKNSLYLQMNSLRDEDTAVYYCAR
		X	DP-52	EDQLVESGGGLVQPGGSLRPSCAASGFAFS	S	YVLH	WVRRAPGKGPEWVS	AIG	TGGDTYYADSVMG	RFTISRDNAKKSLYLQMNSLIAEDMAVYYCAR
	1-3	R2	H11[15]/DP-53	EVQLVESGGGLVQPGGSLRLSCAASGFTFS	S	YWMH	WVRQAPGKGLVWVS	RINS	DGSSTTYADSVKG	RFTISRDNAKNTLYLQMNSLRAEDTAVYYCAR
	1-3	R0	DP-54	EVQLVESGGGLVQPGGSLRLSCAASGFTFS	S	YWMS	WVRQAPGKGLEWVA	NIKQ	DGSEKYYVDSVKG	RFTISRDNAKNSLYLQMNSLRAEDTAVYYCAR
V_H4	1-1	R2	Tou-VH4.21[16]	QVQLQQWGAGLLKPSETLSLTCAVYGGSFS	G	YYWS	WIRQPPGKGLEWIG	EII	HSGSTNYNPSLKS	RVTISVDTSKNQFSLKLSSVTAADTAVYYCAR
	1-1	R0	$V_H5^{10}/V_H4.21^{17}$/DP-63	QVQLQQWGAGLLKPSETLSLTCAVYGGSFS	G	YYWS	WIRQPPGKGLEWIG	EIN	HSGSTNYNPSLKS	RVTISVDTSKNQFSLKLSSVTAADTAVYYCAR
	1-1		V_{58}^{18}	QVQLQQWGAGLLKPSETLSLTCAVYGGSVS	G	YYWS	WIRQPPGKGLEWIG	YIY	YSGSTNNNPSLKS	RATISVDTSKNQFSLNLSSVTAADTAVYYCCAR
	3-1	R1	DP-64	QLQLQESGSGLVKPSQTLSLTCAVSGGSIS	SGGYSWS		WIRQPPGKGLEWIG	YIY	HSGSTYYNPSLKS	RVTISVDRSKNQFSLKLSSVTAADTAVYYCAR
	3-1	R0	DP-65	QVQLQESGPGLVKPSQTLSLTCTVSGGSIS	SGGYYWS		WIRQHPGKGLEWIG	YIY	YSGSTYYNPSLKS	RVTISVDTSKNQFSLKLSSVTAADTAVYYCAR
	3-1	R1	V_{71-2}^{4}/DP-66	QVQLQESGPGLVKPSETLSLTCTVSGGSVS	SGSYYWS		WIRQPPGKGLEWIG	YIY	YSGSTNYNPSLKS	RVTISVDTSKNQFSLKLSSVTAADTAVYYCAR
	2-1	R8	DP-67	QVQLQESGPGLVKPSETLSLTCAVSGYSIS	SG	YYWG	WIRQPPGKGLEWIG	SIY	HSGSTYYNPSLKS	RVTISVDTSKNQFSLKLSSVTAADTAVYYCAR
	2-1		V_H^{SP19}/VH-JA[20]/$V_H4.22^{17}$	QVQLQESGPGLVKPSETLSLTCTVSGYSIS	SG	YYWG	WIRQPPGKGLEWIG	SIY	HSGSTYYNPSLKS	RVTISVDTSKNQFSLKLSSVTAADTAVYYCAR
	2-1	R13	V_{12G-1}^{18}/1.9II[3]/$V_H4.13^{17}$/DP-68	QVQLQESGPGLVKPSDTLSLTCAVSGYSIS	SS	NWWG	WIRQPPGKGLEWIG	YIY	YSGSTYYNPSLKS	RVTISVDTSKNQFSLKLSSVTAVDTAVYYCAR
	2-1		hv4005[21]	QVQLQESGPGLVKPSQTLSLTCAVSGYSIS	SS	NWWG	WIRQPPGKGLEWIG	YIY	YSGSIYYNPSLKS	RVTMSVDTSKNQFSLKLSSVTAVDTAVYYCAR
	2-1	H	V_{11}^{17}	QVQLQESGPGLVKPSETLSLTCVVSGGSIS	SS	NWWS	WVRQPPGKGLEWIG	EIY	HSGNPNYNPSLKS	RVTISIDKSKNQFSLKLSSVTAADTAVYYCAR
	2-1		$V_H4.17^{17}/V_H4.23^{17}$/DP-69	QVQLQESGPGLVKPSETLSLTCVVSGGSIS	SS	NWWS	WVRQPPGKGLEWIG	EIY	HSGSPNYNPSLKS	RVTISVDKSKNQFSLKLSSVTAADTAVYYCAR
	2-1		$V_{79}^{18}/V_H4.19^{17}/V_{IV-4b}^{2}$	QVQLQESGPGLVKPPGTLSLTCAVSGGSIS	SS	NWWS	WVRQPPGKGLEWIG	EIY	HSGSTNYNPSLKS	RVTISVDKSKNQFSLKLSSVTAADTAVYCCAR
	2-1	R0	DP-70	QVQLQESGPGLVKPSGTLSLTCAVSGGSIS	SS	NWWS	WVRQPPGKGLEWIG	EIY	HSGSTNYNPSLKS	RVTISVDKSKNQFSLKLSSVTAADTAVYYCAR
	3-1	R0	$V_H4.18^{17}$	QLQLQESGPGLVKPSETLSLTCTVSGGSIS	SSSYYWG		WIRQPPGKGLEWIG	SIY	YSGSTYYNPSLKS	RVTISVDTSKNQFSLKLSSVTAADTAVYYCAR
	3-1		V_{2-1}^{18}	QLQLQESGPGLVKPSETLSLTCTVSGGSIS	SSSYYWG		WIRQPPGKGLEWIG	SIY	YSGSTYYNPSLKS	RVTISVDTSKNHFSLKLSSVTAADTAVYYCAR
	1-1	R1	V_{IV-4}^{2}	QVQLQESGPGLVKPSETLSLTCTVSGGSIS	S	YYWS	WIRQPAGKGLEWIG	RIY	TSGSTNYNPSLKS	RVTMSVDTSKNQFSLKLSSVTAADTAVYYCAR
	1-1	R0	$V_H4.11^{17}$/DP-71	QVQLQESGPGLVKPSETLSLTCTVSGGSIS	S	YYWS	WIRQPPGKGLEWIG	YIY	YSGSTNYNPSLKS	RVTISVDTSKNQFSLKLSSVTAADTAVYYCAR
	1-1	R5	V_{71-4}^{4}	QVQLQESGPGLVKPSETLSLTCTVSGGSVS	S	YYWS	WIRQPPGKGLEWIG	YIY	YSGSTNYNPSLKS	RVTISVDTSKNQFSLKLSSVTAADTAVYYCAR
	1-1		$V_H4.16^{17}$	QVQLQESGPGLVKPSDTLSLTCTVSGGSIS	S	YYWS	WIRQPPGKGLEWIG	YIY	YSGSTNYNPSLKS	RVTISVDTSKNQFSLKLSSVTAADTAVYYCAR
V_H5	1-2	R0	VH251[22]/DP-73	EVQLVQSGAEVKKPGESLKISCKGSGYSFT	S	YWIG	WVRQMPGKGLEWMG	IIYP	GDSDTRYSPSFQG	QVTISADKSISTAYLQWSSLKASDTAMYYCAR
	1-2		V_HVJB^{17}	EVQLVQSGAEVKKPGESLKISCKGSGYSFT	S	YWIG	WVRQMPGKGLEWMG	IIYP	GDSDTRYSPSFQG	QVTISADKPISTAYLQWSSLKASDTAMYYCAR
	1-2	R6	V_HVCW^{17}	EVQLVQSGAEVKKPGESLKISCKGSGYSFT	S	YWIG	WVRQMPGKGLEWMG	IIYP	GDSDTRYSPSFQG	QVTISADKSISTAYLQWSSLKASDTAMYYCAR
	1-2	H,N	$1-v^{3}$	EVQLLQSAAEVKRPGESLRISCKTSGYSFT	S	YWIH	WVRQMPGKELEWMG	SIYP	GNSDTRYSPSFQG	HVTISADSSSSTAYLQWSSLKASDAAMYYCVR
	1-2		VH32[23]	EVQLVQSGAEVKKPGESLRISCKGSGYSFT	S	YWIS	WVRQMPGKGLEWMG	RIDP	SDSYTNYSPSFQG	HVTISADKSISTAYLQWSSLKASDTAMYYCAR
	1-2	R0	V_HVRG^{17}/V_HVMW^{17}	EVQLVQSGAEVKKPGESLRISCKGSGYSFT	S	YWIS	WVRQMPGKGLEWMG	RIDP	SDSYTNYSPSFQG	HVTISADKSISTAYLQWSSLKASDTAMYYCAR
V_H6	3-5	R0	V_H-VI[24]/6-1G1[3]/DP-74	QVQLQQSGPGLVKPSQTLSLTCAISGDSVS	SNSAAWN		WIRQSPSRGLEWLG	RTYYR	SKWYNDYAVSVKS	RITINPDTSKNQFSLQLNSVTPEDTAVYYCAR

Fig. 2(a) *continued*

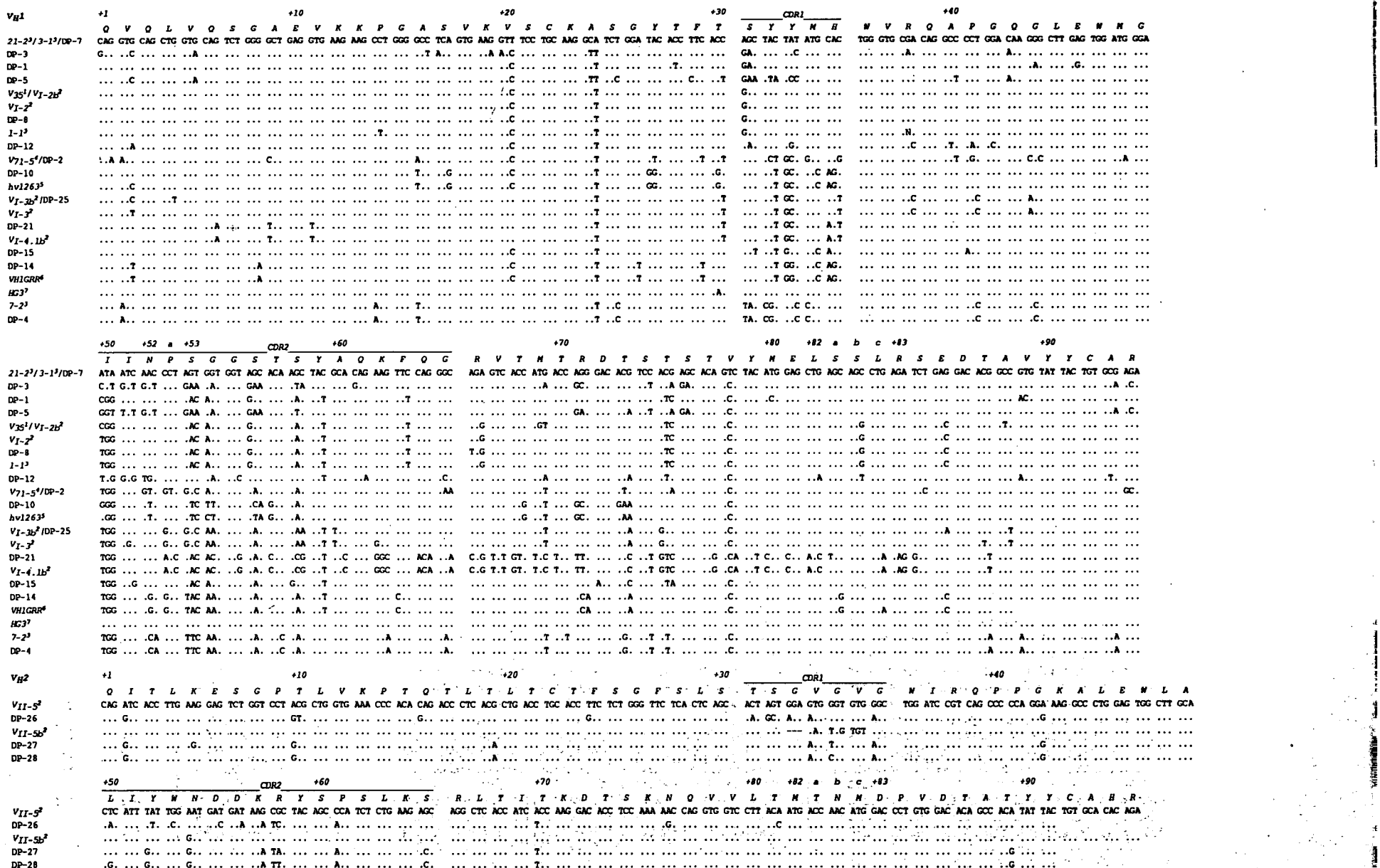


Fig. 2(b)

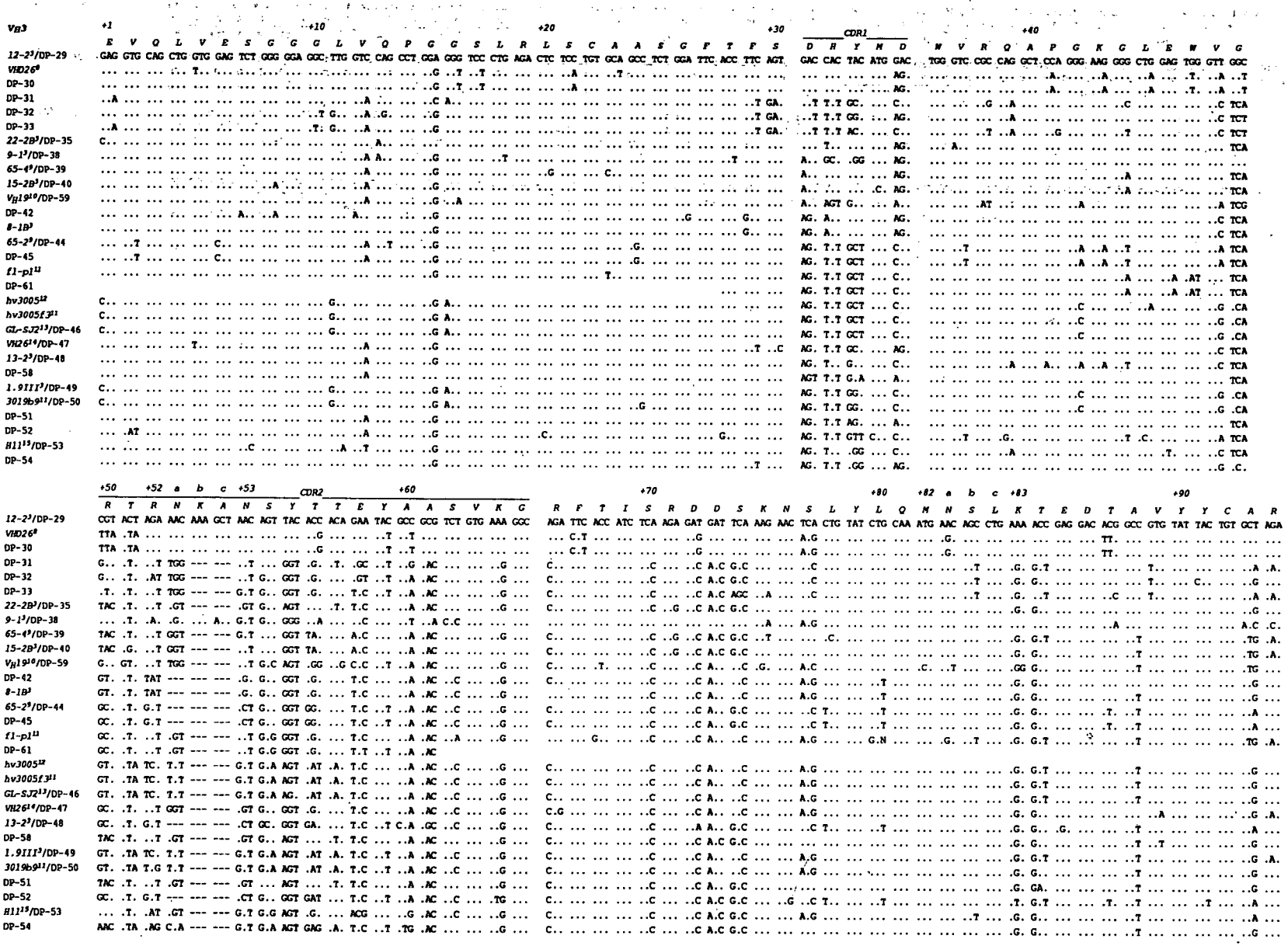


Fig. 2(b) *continued*

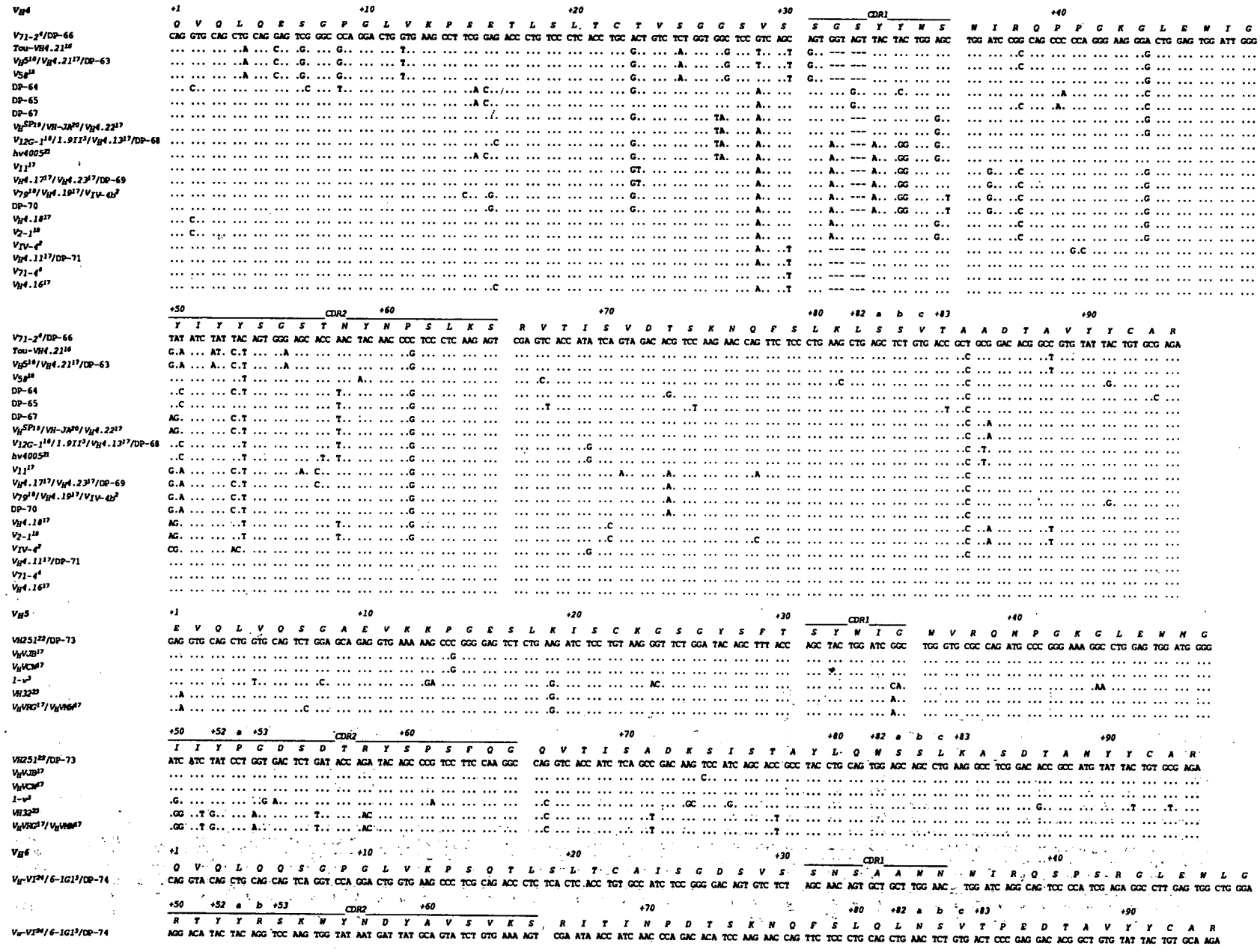


Fig. 2(b) *continued*

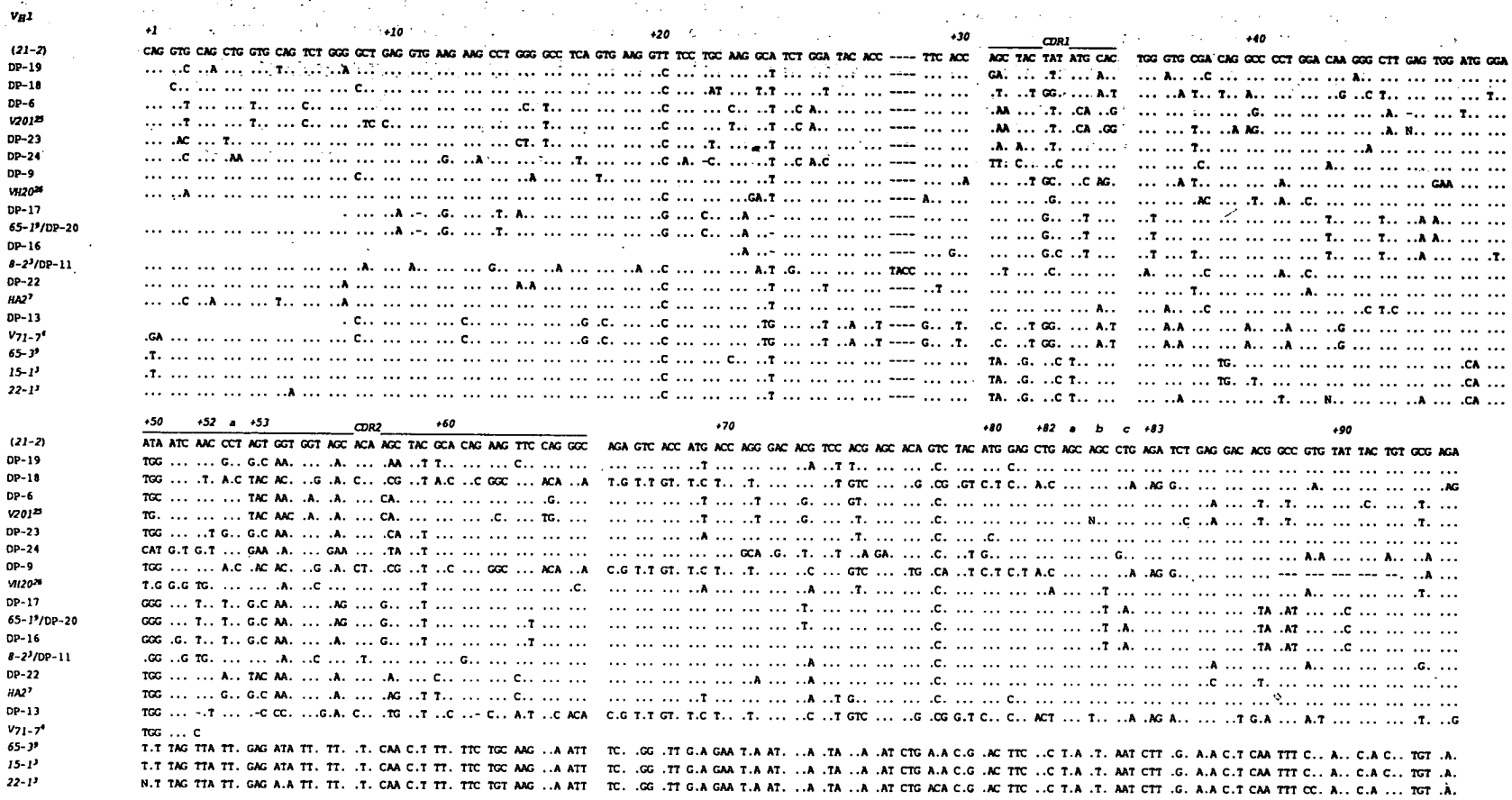

VH1
+1 +10 +20 +30
CDR1
+40
(21-2) CAG GTG CAG CTG GTG CAG TCT GGG GCT GAG GTG AAG AAG CCT GGG GCC TCA GTG AAG GTT TCC TGC AAG GCA TCT GGA TAC ACC ---- TTC ACC AGC TAC TAT ATG CAC TGG GTG CGA CAG GCC CCT GGA CAA GGG CTT GAG TGG ATG GGA
DP-19
DP-18
DP-6
VI201[25]
DP-23
DP-24
DP-9
VH120[26]
DP-17
65-1[9]/DP-20
DP-16
8-2[3]/DP-11
DP-22
HA2[7]
DP-13
V71-7[4]
65-3[9]
15-1[3]
22-1[3]
+50 +52 a +53 CDR2 +60 +70 +80 +82 a b c +83 +90
(21-2) ATA ATC AAC CCT AGT GGT GGT AGC ACA AGC TAC GCA CAG AAG TTC CAG GGC AGA GTC ACC ATG ACC AGG GAC ACG TCC ACG AGC ACA GTC TAC ATG GAG CTG AGC AGC CTG AGA TCT GAG GAC ACG GCC GTG TAT TAC TGT GCG AGA
DP-19
DP-18
DP-6
VI201[25]
DP-23
DP-24
DP-9
VH120[26]
DP-17
65-1[9]/DP-20
DP-16
8-2[3]/DP-11
DP-22
HA2[7]
DP-13
V71-7[4]
65-3[9]
15-1[3]
22-1[3]

Fig. 2(c)

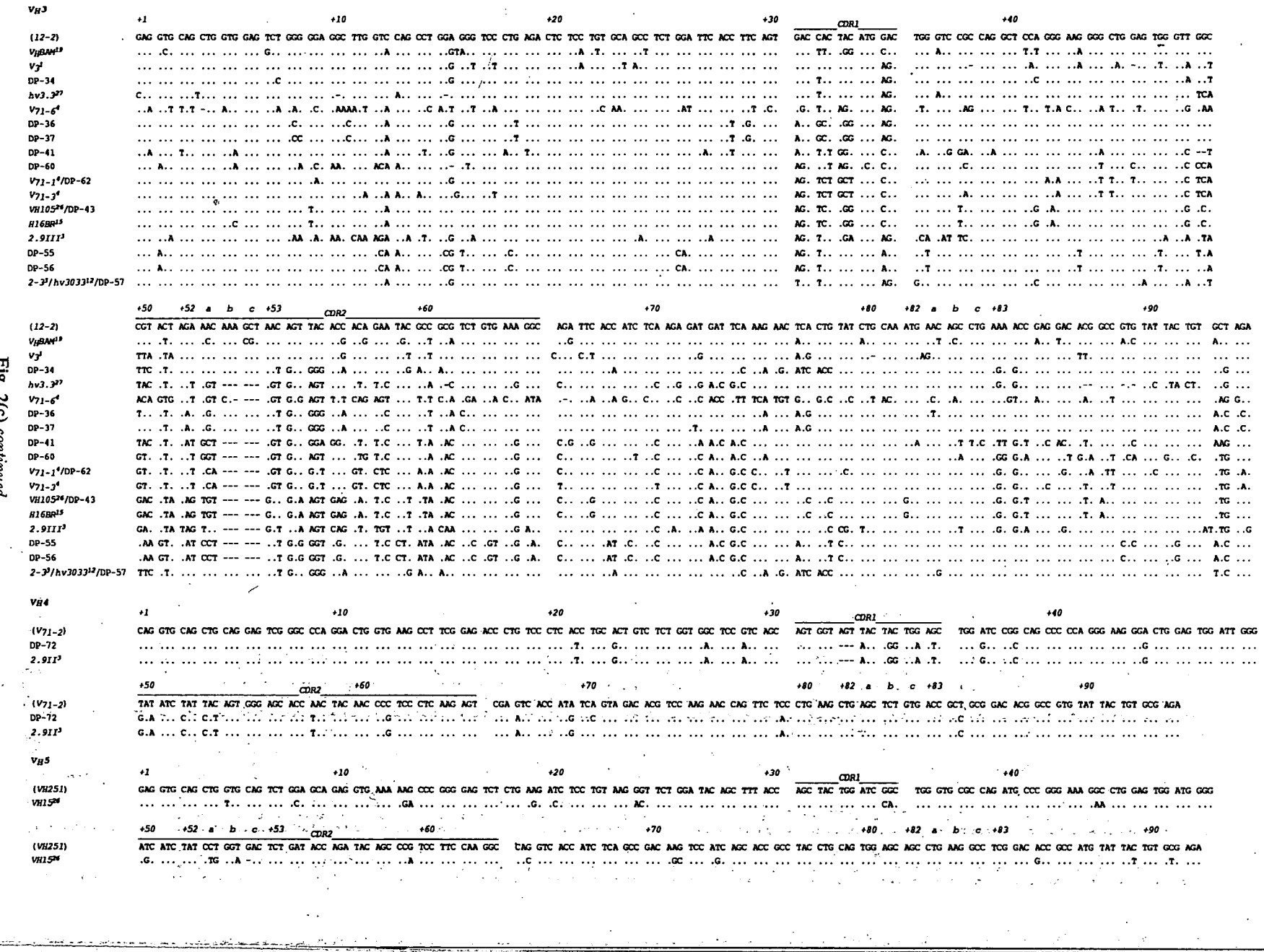
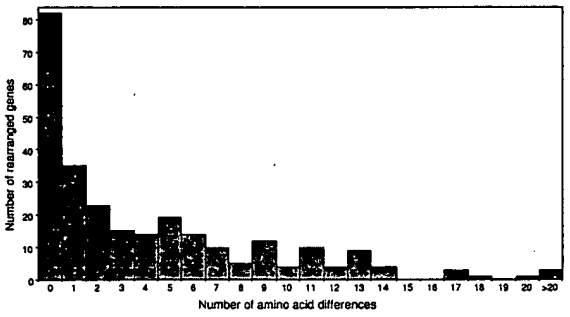


Fig. 2(c) *continued*

Figure 2. Directory of germline V_H segments. Genes are divided into their respective families and the framework (FR) and complementarity-determining regions (CDR) are as defined by Kabat *et al.* (1991). Where 2 genes have identical nucleotide sequences, both are shown separated by a slash. Genes prefixed DP are from this study. Previously published genes are shown in italics and suffixed according to source: [1]Matsuda *et al.* (1988); [2]Shin *et al.* (1991); [3]Berman *et al.* (1988); [4]Kodaira *et al.* (1986); [5]Chen *et al.* (1989); [6]Friedman *et al.* (1991); [7]Rechavi *et al.* (1983); [8]Buluwella *et al.* (1988); [9]Matsuda *et al.* (1990); [10]Baer *et al.* (1988); [11]Olee *et al.* (1991); [12]Chen (1990); [13]Pascual *et al.* (1990); [14]Mathyssens & Rabbitts (1980), corrected by Chen *et al.* (1988); [15]Rechavi *et al.* (1982); [16]van Es *et al.* (1991); [17]Sanz *et al.* (1989*c*); [18]Lee *et al.* (1987); [19]Baer *et al.* (1985); [20]Denny *et al.* (1986); [21]Chen & Yang (1990); [22]Shen *et al.* (1987), corrected by Sanz *et al.* (1989*c*); [23]Humphries *et al.* (1988), corrected by Sanz *et al.* (1989*c*); [24]Buluwela & Rabbitts (1988); [25]Takahashi *et al.* (1984); [26]Humphries *et al.* (1988); [27]Turnbull *et al.* (1987). A number of published V_H segments have discrepancies between the sequence described in the original literature and that submitted to GenBank. We have used the sequence from the literature. (a) Protein sequences of V_H segments with open reading frames. Sequences are shown in single-letter amino acid code and have been aligned according to Kabat *et al.* (1991), except in CDR1, where padding is according to the H1 loop structure (Chothia *et al.*, 1992). Sequences are arranged alphabetically by CDR1 and, where these are identical, by CDR2. Sequences that have identical translated CDRs 1 and 2 are grouped with brackets. The canonical structure classes of H1 (CDR1) and H2 (CDR2) are shown, separated by a dash (see Chothia *et al.*, 1992). The sequence marked " . " has insufficient data to be classified. Features: *X*, may be defective on structural grounds (see Chothia *et al.*, 1992); *H*, heptamer differs from the conserved 5′-CACAGTG-3′ motif; *N*, nonamer sequence differs from the family consensus; *R*, seen as rearranged gene (see Fig. 3 and Results), the smallest number of amino acid differences between each rearranged gene and its closest germline counterpart is shown. Thus, $V_{I\text{-}2}$ and DP-8 have identical encoded CDRs 1 and 2, have H1 canonical structure class 1, H2 canonical structure class 3 and have both been seen as rearranged genes with no amino acid differences. (b) DNA sequences of V_H segments with open reading frames. Sequences have been aligned to a master sequence (see Results) and nucleotides identical to this sequence are shown as dots and deletions are indicated by a dash. Numbering is according to the corresponding amino acid residue. (c) V_H pseudogenes. Sequences of germline V_H segments with either frame shifts or stop codons, which are therefore considered to be pseudogenes. Because of their different lengths and heterogeneity, they have been aligned to the master sequence (which for each family is the same as in (b), above) by a maximum alignment program (MacVector). Nucleotides that are identical to the master sequence are shown as dots and deletions are indicated by a dash. Insertions have been placed in between adjacent codons.

Germline Segment	Rearranged Gene	Reference	Number of Amino Acid Changes
DP-3	S1P19	(1)	5
V_H-2	16-4	(2)	0
DP-8	15-4	(2)	0
DP-10	AND	(3)	0
bv1263	III-2R	(4)	3
V_H-3b	VEDD10	(5)	0
V_H-4.1b	RF-TS3	(6)	1
DP-15	216E	(7)	2
DP-14	12B	(7)	1
VH1GRR	LS7	(8)	4
21-2	MO30	(9)	3
DP-26	26E	(7)	0
DP-27	M60	(10)	3
DP-28	26D	(7)	0
V_H-5	26A	(7)	0
DP-31	C6H	(11)	0
DP-32	6M9	(12)	2
DP-33	6A1	(13)	4
22-2B	H2F	(6)	1
9-1	M26	(10)	0
DP-42	60P2	(14)	1
8-1B	14J	(7)	1
f1-p1	12H	(7)	10
bv3005f3	M72	(10)	0
GL-5J2	A39	(15)	0
VH26	12C	(7)	0
13-2	36-1	(2)	0
DP-58	21SD	(7)	0
1.9111	αBSA3	(16)	0
3019b9	5A10	(17)	0
DP-51	11G9-10B4	(15)	3
H11	21SL	(7)	2
DP-54	X31	(18)	0
Tou-VH4.21	Fom-1	(19)	2
V_H5	21SH	(7)	0
DP-64	21SG	(7)	1
DP-65	14L	(7)	0
V_H1-2	216H	(7)	1
DP-67	A45S	(20)	8
V_H12G-1	Ab26	(21)	13
DP-70	21SC	(7)	0
V_H4.18	14D	(7)	0
V_HV-4	21SA	(7)	1
V_H4.11	7-2	(22)	0
V_H1-4	Pag-1	(23)	5
VH251	28-3	(2)	0
V_HVCW	II-1	(4)	6
V_HVRG	M13	(20)	0
V_H-VI	17-2	(2)	0

(a)



(b)

Figure 3. Assignment of rearranged human V_H genes to their closest germline counterparts. (a) Germline V_H segments and the closest rearranged V_H gene, references are (1) Bridges *et al.* (1991)†; (2) Deane & Norton (1990)‡§; (3) Kipps *et al.* (1989)§; (4) Manheimer-Lory *et al.* (1991)†; (5) Noma *et al.* (1984); (6) Pascual *et al.* (1990)†; (7) Marks *et al.* (1991*b*)‡; (8) Silberstein *et al.* (1989)†; (9) Larrick *et al.* (1989*a*); (10) Schroeder & Wang (1990)¶; (11) Ermel *et al.* (1991)†; (12) Karr *et al.* (1991)¶; (13) Brown *et al.* (1991)†; (14) Schroeder *et al.* (1987)¶; (15) Geng *et al.* (1991)†; (16) Marks *et al.* (1991*a*)‡; (17) see Olee *et al.* (1991)¶; (18) Timmers *et al.* (1991); (19) Bye *et al.* (1992)‡; (20) Schutte *et al.* (1991)¶†§; (21) Sanz *et al.* (1989*a*)†; (22) Desai *et al.* (1990)§; (23) Hughes-Jones *et al.* (1990). (b) Distribution of the number of amino acid differences between each rearranged V_H gene (268 examples) and its closest germline counterpart. Data were taken from the above references and Kenten *et al.* (1982); Takahashi *et al.* (1984); Kudo *et al.* (1985); Mensink *et al.* (1986); Dersimonian *et al.* (1987)†; Shen *et al.* (1987)§; Berman *et al.* (1988); Meeker *et al.* (1988)§; Newkirk *et al.* (1988); Cairns *et al.* (1989)†; Carroll *et al.* (1989); Chen *et al.* (1989)§; Dersimonian *et al.* (1989)†; Gillies *et al.* (1989); Kishimoto *et al.* (1989); Larrick *et al.* (1989*b*); Logtenberg *et al.* (1989)¶†; Nakatani *et al.* (1989); Nickerson *et al.* (1989)¶; Sanz *et al.* (1989*b*)†; Yasui *et al.* (1989); Akahori *et al.* (1990); Felgenhauer *et al.* (1990); Friedlander *et al.* (1990); Guillaume *et al.* (1990)¶†; Robbins *et al.* (1990)†; Roudier *et al.* (1990)†§; Siminovitch & Chen (1990)†; Spatz *et al.* (1990)§; van der Heijden *et al.* (1990); White *et al.* (1990); Andris *et al.* (1990); Ezaki *et al.* (1991)†; Friedman *et al.* (1991)†; Kuppers *et al.* (1991)§; Mortari *et al.* (1991); Pascual *et al.* (1991); Rioux *et al.* (1991)†; Silberstein *et al.* (1991); van Es *et al.* (1991)†; Mierau *et al.* (1992) . Some of the references include sequences from several different rearranged V_H genes: all of the sequences (except where the genes could not be assigned, see Results) have been used. For key to annotation of references (†, ‡, § and ¶) see Discussion.

family:

(V_H1, 5′-TCAGAAACC-3′;
V_H2, 5′-ACAAAAACC-3′;
V_H3, 5′-ACACAAACC-3′;
V_H4, 5′-ACAAAAACC-3′ or
5′-ACACAAACC-3′;
V_H5, 5′-TCTAAAACC-3′;
V_H6, 5′-ACACAAACC-3′).

Where the nonamer sequence differs from the family consensus the V_H segment is marked *N*.

We compiled a database of 292 rearranged (but not necessarily functional) V_H genes and assigned 268 of these, from 64 different sources (see legend to Fig. 3), to their closest germline counterparts. In Figure 3(a) we list the V_H segments, each with an example of a rearranged V_H gene with the smallest number of amino acid differences. These data are summarized in Figure 2(a), with sequences marked *R* having rearranged counterparts with the indicated number of amino acid differences. The distribution of the number of amino acid differences across all 268 assigned rearranged genes is shown in Figure 3(b): 215 of the 292 rearranged V_H genes in our database have germline counterparts seen in DP (data not shown).

We were unable to assign 24 rearranged genes from the V_H3 (VDJ191, Mensink *et al.* (1986); X51, X61, X71, Timmers *et al.* (1991); K6H6, K4B8, K5B8, K5G5, K6F5, K5C7, Kon *et al.* (1987); 333, 1H1, 2C12, 2A12, 1B11, 112, 115 and 126, Cleary *et al.* (1986)) and V_H4 (TS2, Shen *et al.* (1987); HIVB, Andris *et al.* (1991); C6B2, Hoch & Schwaber (1987); 2A4, Davidson *et al.* (1990); 12-3, 30-2, Deane & Norton (1990)) families. Of these, 12-3 (Deane & Norton, 1990) is almost certainly the result of a PCR cross-over and the others appear to be derived from a possible four to six unknown germline V_H segments.

(e) *Germline sequence variability*

Based on data from Figure 2(a), we have constructed variability plots, shown in Figure 4, for germline V_H segments with open reading frames from all six families, as well as separate plots for the V_H1 and V_H3 families. We only excluded those sequences marked *X* which may be defective on structural grounds (see above). At each position, a variability score was calculated as the number of different amino acids at that position, divided by the percentage frequency of occurrence of the most common amino acid (see Kabat *et al.*, 1991).

4. Discussion

(a) *Cloning and sequencing strategy*

Our strategy for sequencing V_H segments by PCR amplification of genomic DNA is based on the use of

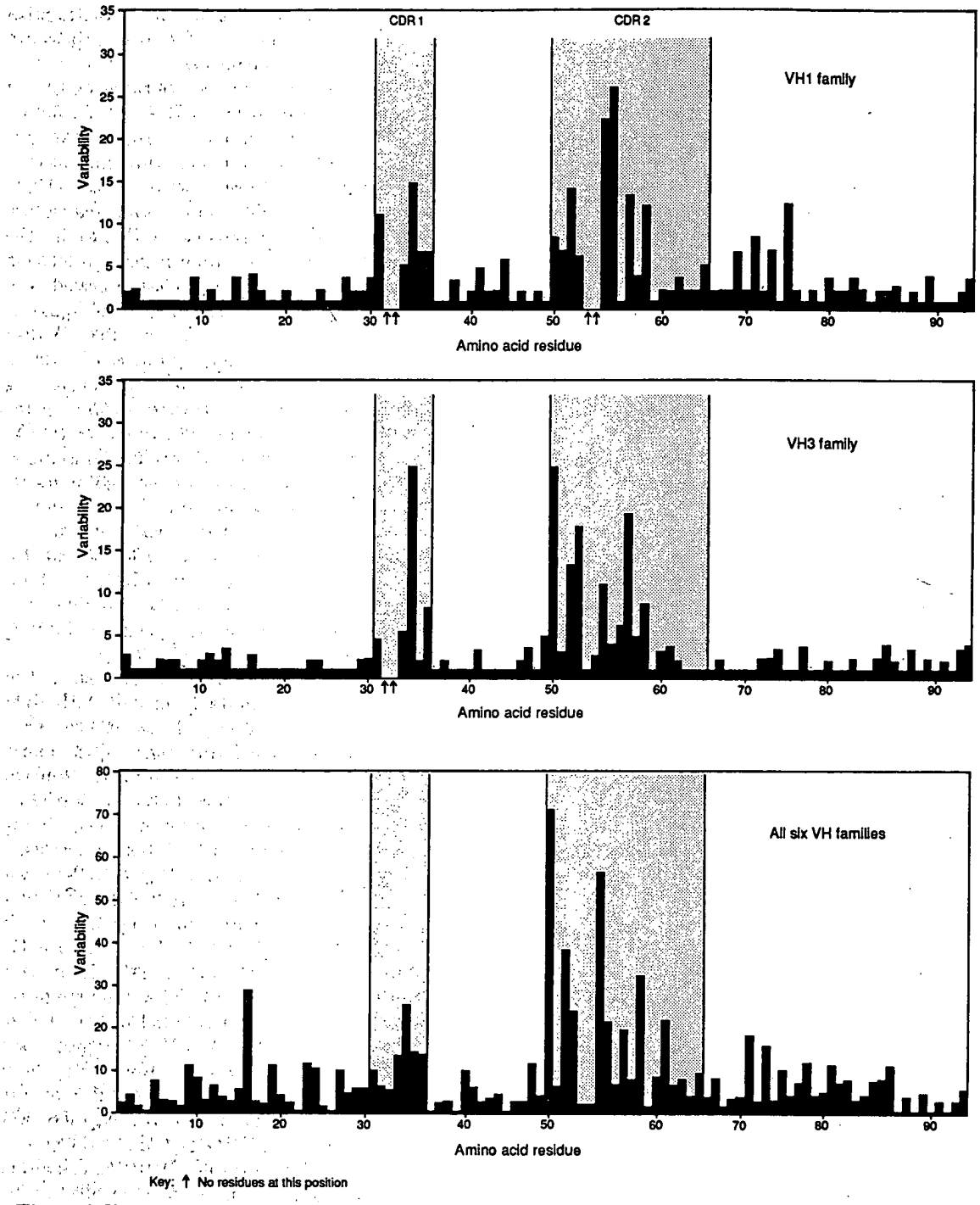


Figure 4. Variability plot for germline V_H segments. Variability was calculated (see Results) across protein sequences shown in Fig. 2(a), but excluding those that are likely to be defective on structural grounds (marked *X*). Plots were produced for the V_H1 family, V_H3 family and across all 6 families.

family-specific primers designed from the sequences of the six *known* V_H families. We were able to assign most of the rearranged V_H genes to germline V_H segments in Figure 2 with few differences in amino acid sequences (Fig. 3(b)), but may have missed V_H segments that are significantly different in the primer regions; for example, we did not find the germline counterparts of the rearranged genes determined by Cleary *et al.* (1986). Indeed, they have been classified as belonging to a new family (V_H7) by some authors (Schroeder *et al.*, 1990), but they might also be highly mutated genes derived from a known germline V_H segment (especially as they were derived from B-cell lymphomas).

Since our aim was to determine the structural repertoire of human V_H segments, the majority of primers were designed to amplify genes with "functional" heptamer recombination sequences

(5′-CACAGTG-3′). We have therefore missed some genes with different heptamers, which presumably includes some pseudogenes. For example, three sequences which were amplified with internal V_H3 primers and have unusual heptamer sequences, DP-59/V_H19, DP-60 and DP-62/V_{71-1}, were not amplified using the heptamer primers. It is, however, unclear what constitutes a functional heptamer; indeed, in a recent study, Shin *et al.* (1991) discovered two V_H2 segments with an unusual heptamer sequence (5′-CACAAAG-3′). One of these segments has been seen as a rearranged gene (see Fig. 3(a)). This suggests that the 5′-CACAGTG-3′ heptamer sequence is not the only one used for recombination and, consequently, that the D segment heptamer may also be degenerate. This, and the fact that these V_H2 segments (V_{11-5}) have an additional amino acid residue in framework 3, may explain the poor performance of our V_H2 primers and the relatively low number of V_H2 segments isolated here.

In addition, those genes with open reading frames (Fig. 2(a) and (b)) may be non-functional for other reasons. For example, the V_H1 segments 1-1 (Berman *et al.*, 1988) and V_{71-5} (Kodaira *et al.*, 1986) have single base differences in the recombination nonamer and the leader intron splice site, respectively, and 1-v (Berman *et al.*, 1988) has a frame shift in the leader exon. Certain V_H segments may also be defective on structural grounds (marked *X* in Fig. 2(a), see Chothia *et al.*, 1992).

To avoid polymerase copying errors, we screened more than 2000 clones using motif- or gene-specific oligonucleotide probes to ensure identical nucleotide sequences from *two independent* amplifications. Copying errors fell into two categories: base substitutions and PCR cross-overs. Substitutions might have been reduced by using a polymerase with a 5′ to 3′ proof-reading activity such as Vent (New England Biolabs, Beverly, MA) or Pfu (Stratagene, La Jolla, CA) DNA polymerases. However, under a range of conditions, these polymerases performed poorly (data not shown). PCR cross-overs occurred within the region of greatest homology, and were most easily detected by unexpected combinations of CDR1 and CDR2 due to a cross-over in framework 2. This emphasizes the importance of confirmation from independent amplifications rather than from multiple clones of the same PCR; indeed, germline V_H segments hv3005b54, hv3019b13, hv3019b18 (Olee *et al.*, 1991) and $V_H4.12$, $V_H4.14$, $V_H4.15$ (Sanz *et al.*, 1989*c*) may be the result of PCR artifacts (see above).

(b) *Polymorphism*

In our directory (Fig. 2), which contains data from many individuals, we have a total of 122 V_H segments with different nucleotide sequences (41 V_H1 segments, 5 V_H2 segments, 46 V_H3 segments, 22 V_H4 segments, 7 V_H5 segments and 1 V_H6 segment), including 83 V_H segments with open reading frames and 39 pseudogenes. However, we cannot exclude polymorphism and allelic variation or distinguish between identical V_H genes at different loci (possibly the result of a recent duplication).

Southern blot analyses of restriction digests of DNA using cDNA probes (van Dijk *et al.*, 1991), germline coding and flanking region probes (Souroujon *et al.*, 1989) or short sequence-specific probes (Sanz *et al.*, 1989*c*; Sasso *et al.*, 1990; van Dijk *et al.*, 1991) have demonstrated restriction fragment length polymorphisms (RFLPs) in the V_H3, V_H4 and V_H5 families. Some insertion/deletion polymorphisms have also been characterized and shown to involve, for example, at least one V_H2, one V_H3 and one V_H5 gene (Chen & Yang, 1990; Walter *et al.*, 1990), and one V_H1 gene (Shin *et al.*, 1991). Indeed, we failed to clone from DP several V_H segments reported in the literature, despite using suitable PCR primers and probes. Some of the V_H segments not amplified from DP are also missing in other individuals. For example, of the V_H4 segments not amplified from our donor, one (V_{58}) seen in a Japanese study (Lee *et al.*, 1987) was not found in an American study (Sanz *et al.*, 1989*c*) and the absence of a second V_H5 segment, VH32 (see Sanz *et al.*, 1989*c*), from our donor may be due to a deletion polymorphism affecting V_H5 genes in 50% of individuals (Sam *et al.*, 1988).

In our directory, we found that the nucleotide sequences of 23 V_H segments from DP with open reading frames were identical to those from unrelated individuals. We found other V_H segments with a few nucleotide differences but with identical translated CDRs 1 and 2 (bracketed in Fig. 2(a)) and these may correspond to different alleles. Thus, the following V_H1 segments differ by one nucleotide: V_{I-2}, DP-8 and 1-1; DP-21 and $V_{I-4.1b}$; DP-14 and VH1GRR; 21-2/3-1/DP-7 and HG3; 7-2 and DP-4. The following V_H3 segments differ by one to six nucleotides: VHD26 and DP-30; DP-42 and 8-1B; 65-2/DP-44 and DP-45; f1-p1 and DP-61; hv3005, hv3005f3 and GL-SJ2/DP-46. The following V_H4 segments differ by one or two nucleotides: DP-67 and V_H^{SP}/VH-JA/$V_H4.22$; V_{79}/$V_H4.19$/V_{IV-4b} and DP-70; $V_H4.18$ and V_{2-1}; $V_H4.11$/DP-71, V_{71-4} and $V_H4.16$. The following V_H5 segments differ by one or two nucleotides: VH251/DP-73, V_HVJB and V_HVCW; VH32 and V_HVRG/V_HVMW. Of course, other V_H segments, for example, DP-10 and hv1263, and V_{I-3b}/DP-25 and V_{I-3} may also be alleles, but they encode differences in the CDRs and have therefore been grouped separately. This is consistent with the suggestion that even diverse V_H segments (V_{II-5} and V_{II-5b}; V_{IV-4} and V_{IV-4b}) could be alleles (Shin *et al.*, 1991).

Hence, we find a "core" of V_H segments with open reading frames that are highly conserved in the antigen binding regions and differ by only a small number of nucleotides in the framework regions. This limited sequence polymorphism between unrelated individuals together with the insertion/deletion polymorphism agrees with the suggestion that the germline V_H repertoire is derived from a population of diverse haplotypes with a small

number of alleles at each locus (Sasso *et al.*, 1990; van Dijk *et al.*, 1991).

In contrast to the limited sequence polymorphism in V_H segments with open reading frames, only five pseudogenes amplified from DP are identical to V_H segments seen in unrelated individuals and a further five pairs differ by one or two nucleotides. The finding that certain pseudogenes are identical, or are very similar, in unrelated individuals (see Fig. 2(c)) has been previously noted (Kodaira *et al.*, 1986) and may indicate a physiological role for them, possibly as donors for gene conversion, as in the chicken (Reynaud *et al.*, 1989).

(c) *Assignment of rearranged genes*

As shown in Figure 3(b), the majority of rearranged genes, usually derived from mRNA, are very closely related to their germline counterparts. This confirms that these germline genes can be rearranged and transcribed and are probably translated into protein. Some of the differences between the rearranged and germline genes could be due to germline polymorphism, but as this is limited (see above), the majority are probably caused by somatic mutation. In a few examples, the sequences of the rearranged V_H genes appear to be a composite of two V_H segments (215B and 216G; Marks *et al.*, 1991*b*), which presumably arose by PCR cross-over.

The assignment of rearranged human V_H genes to their germline counterparts may help in dissecting mechanisms of the human immune system. It enables us to determine the relative usage of particular V_H segments (the possible underexpression of V_H1 segments and overexpression of V_H4 segments) and the number and location of somatic mutations by which a particular antibody has been shaped. It also allows us to differentiate between immune responses that utilize V_H segments with different levels of somatic mutation. For example, it has been repeatedly suggested that foetal antibodies and autoantibodies are dominated by rarely mutated or unmutated germline V_H genes and that these antibodies are often polyreactive (see Chen *et al.*, 1990; Hillson & Perlmutter, 1990; Siminovitch & Chen, 1990; Pascual & Capra, 1991).

Using our database of human rearranged V_H genes we find that about three-quarters of the genes of foetal origin are germline at the level of amino acid sequence and the rest have no more than five amino acid changes (see references marked ¶ in Fig. 3 legend). However, in the case of autoantibodies (autoimmunity related V_H genes, see references marked † in Fig. 3 legend) there is no clear difference in the overall number of amino acid changes compared to rearranged V_H genes found in normal peripheral blood lymphycytes (see references marked ‡ in Fig. 3 legend). This does not support the concept that autoantibodies are mainly encoded by rarely mutated or unmutated V_H genes and reflects the current uncertainty about the origin of autoantibodies and the role of antigen stimulation (Dersimonian *et al.*, 1990). Other interesting features emerge for different B cell malignancies. Whereas most of the V_H genes isolated from acute lymphoblastic leukaemia (ALL) patients are rarely mutated or unmutated (Berman *et al.*, 1988; Carroll *et al.*, 1989; Deane & Norton, 1990), about half the V_H genes isolated from patients with chronic lymphocytic leukaemia (CLL) contain more than six amino acid changes (see references marked § in Fig. 3 legend). Very highly mutated V_H genes (17, 20, 43 amino acid changes) have been detected in other B cell tumours, such as myelomas (White *et al.*, 1990; Kenten *et al.*, 1982; Yasui *et al.*, 1989).

(d) *Number of human V_H segments*

Estimates of the number of human V_H segments per individual have been based on restriction digests of genomic DNA probed for each V_H family, but are likely to be underestimates (due to bands co-migrating on the gel). For example, Southern blot analyses of digested DNA from HeLa and LA-N-5 cell lines yielded 60 to 80 hybridizing fragments (Berman *et al.*, 1988) but the authors estimated the total number of V_H segments to be between 100 and 200. More recently, two-dimensional pulse field gel electrophoresis of digested homozygous DNA (Walter *et al.*, 1990) suggested a total of 76 V_H segments (25 V_H1 segments, 5 V_H2 segments, 28 V_H3 segments, 14 V_H4 segments, 3 V_H5 segments and 1 V_H6 segment).

We have cloned and sequenced 74 human V_H segments (25 V_H1 segments, 3 V_H2 segments, 34 V_H3 segments, 10 V_H4 segments, 1 V_H5 segment and 1 V_H6 segment). Fifty-one of these have open reading frames, and 23 contain either frame shifts or stop codons and are therefore considered to be pseudogenes. While the number of pseudogenes amplified from DP is likely to be an underestimate due to primer bias, the number of V_H segments with open reading frames (51) seems to correspond to the coding repertoire of an individual. Indeed, 215 of 292 rearranged V_H genes from different (non-DP) individuals have germline counterparts seen in DP. The extent to which our individual is representative of the human population as a whole depends on the *exact* nature of polymorphism within the V_H locus. To determine this, we need a physical map of the V_H segments from individuals with different genetic backgrounds, in which individual V_H loci have been sequenced. This would tell us the number of different sequences in the human V_H segment pool, the total number of loci and the number of alleles at each locus.

(e) *Structural diversity of human germline V_H segments*

In order to focus on the structural diversity of antigen binding sites implicit in the germline V_H repertoire of the human population, we grouped together (bracketed in Fig. 2(a)) those V_H segments that encode identical CDRs 1 and 2. We have selected those V_H segments with rearranged counter-

parts (marked *R* in Fig. 2(a)) and excluded a few V_H segments (marked *X* in Fig. 2(a)), which appear to be defective on structural grounds (Chothia *et al.*, 1992) and therefore are unlikely to contribute to the functional V_H repertoire.

This suggests that the structural diversity encoded by human germline V_H segments is determined by a minimum of 43 groups of rearranged V_H segments, each encoding identical CDR loops. This figure is likely to increase as rearranged counterparts of other germline V_H segments in Figure 2(a) are discovered and as a few additional germline segments are determined from different individuals. However, those V_H segments with heptamers other than the 5′-CACAGTG-3′ motif (marked *H* in Fig. 2(a)) and those with nonamers that differ from the family consensus (marked *N* in Fig. 2(a)) may be unable to recombine and hence not be expressed.

In order to determine the possible extent of sequence diversity, our variability plots (Fig. 4) are calculated using sequence data from all germline V_H segments of the 43 structural groups and those germline sequences for which no rearranged counterparts have yet been discovered. The use of germline V_H segments eliminates the effects of somatic mutation and sampling bias present in variability plots of rearranged V_H genes (Kabat *et al.*, 1991).

The plots are consistent with the classification of framework (FR) and complementarity-determining regions (CDR) defined by Kabat *et al.* (1991), but new features do emerge. Firstly, variability is higher in CDR2 than in CDR1. Secondly, the hypervariable region of CDR2 only comprises residues 50 to 58, rather than 50 to 65, with the last seven residues of CDR2 (59 to 65) being highly conserved within each of the six families. Thirdly, in addition to CDR1 and CDR2, we find two regions of unusually high variability across all six families. One of them is residue 16 and the other is centred around residue 73 and corresponds to a loop adjacent to CDR2. The region in framework 3 is particularly variable in the V_H1 family and may function by altering the conformation of CDR2 for antigen binding, or make additional contacts directly with the antigen (like in the case of the light chain FR3 in the D1.3/E255 complex: Bentley *et al.*, 1990). Alternatively, it may interact with an unidentified ligand involved in the biology of the B cell response, for example, a superantigen (Schroeder *et al.*, 1990; Sasso *et al.*, 1991).

(f) *Conclusion*

Our strategy has enabled us to determine the human germline V_H segments with open reading frames from a single individual (DP). The comparison with germline V_H segments from other individuals and with 292 rearranged V_H genes suggests that sequence polymorphism is limited, and that the directory could be used to map the V_H locus in different individuals, to determine the usage of specific V_H segments in immune responses and to detect somatic mutation or gene conversion events *in vivo*.

The directory indicates that the structural diversity of the germline repertoire for antigen binding is fixed by about 50 groups of V_H segments. Each group encodes identical hypervariable loops and has been seen as a rearranged gene. The limited diversity encoded by germline V_H segments emphasizes the importance of the additional diversity provided by the D and J_H segments and by somatic mutation. It suggests that our repertoire of V_H segments from DP should be sufficient for building libraries of human antibodies with known components (Winter & Milstein, 1991; Marks *et al.*, 1991*a*).

I.M.T. and G. Walter are supported by the Medical Research Council Human Genome Mapping Project, J.D.M. by the Medical Research Council Aids Directed Programme and M.B.L. by a Medical Research Council/Celltech Ltd. training fellowship. We thank D. J. Perry for his many blood donations and his help with genomic DNA preparation.

References

Akahori, Y., Kurosawa, Y., Kamachi, Y., Torii, S. & Matsuoka, H. (1990). Presence of immunoglobulin (Ig) M and IgG double isotype-bearing cells and defect of switch recombination in hyper IgM immunodeficiency. *J. Clin. Invest.* **85**, 1722–1727.

Alzari, P. M., Spinelli, S., Mariuzza, R. A., Boulot, G., Poljak, R. J., Jarvis, J. M. & Milstein, C. (1990). Three-dimensional structure determination of an anti-2-phenyloxazolone antibody: the role of somatic mutation and heavy/light chain pairing in the maturation of an immune response. *EMBO J.* **9**, 3807–3814.

Amit, A. G., Mariuzza, R. A., Phillips, S. E. & Poljak, R. J. (1986). Three-dimensional structure of an antigen–antibody complex at 2·8 Å resolution. *Science*, **233**, 747–753.

Andris, J. S., Johnson, S., Zolla, P. S. & Capra, J. D. (1991). Molecular characterization of five human anti-human immunodeficiency virus type 1 antibody heavy chains reveals extensive somatic mutation typical of an antigen-driven immune response. *Proc. Nat. Acad. Sci., U.S.A.* **88**, 7783–7787.

Baer, R., Chen, K. C., Smith, S. D. & Rabbitts, T. H. (1985). Fusion of an immunoglobulin variable gene and a T cell receptor constant gene in the chromosome 14 inversion associated with T cell tumors. *Cell*, **43**, 705–713.

Baer, R., Forster, A., Lavenir, I. & Rabbitts, T. H. (1988). Immunoglobulin VH genes are transcribed by T cells in association with a new 5′ exon. *J. Exp. Med.* **167**, 2011–2016.

Bentley, G. A., Boulot, G., Riottot, M. M. & Poljak, R. J. (1990). Three-dimensional structure of an idiotope-anti-idiotope complex. *Nature (London)*, **348**, 254–257.

Berek, C. & Milstein, C. (1988). The dynamic nature of the antibody repertoire. *Immunol. Rev.* **105**, 5–26.

Berman, J. E., Mellis, S. J., Pollock, R., Smith, C. L., Suh, H., Heinke, B., Kowal, C., Surti, U., Cantor, C. R. & Alt, F. W. (1988). Content and organization of the human Ig VH locus: definition of three new

VH families and linkage to the Ig CH locus. *EMBO J.* **7**, 727–738.

Borghesi-Nicoletti, C. & Schulze, D. H. (1991). Polymerase chain reaction of genes flanked by short noncontiguous sequence motifs. *Anal. Biochem.* **192**, 449–452.

Bridges, S. L., Lee, S. L., Koopman, W. J. & Schroeder, H. W. (1991). V_H1 gene segments expressed in rheumatoid synovium. *American College of Rheumatology 55th Annual Scientific Meeting,* 17–21 November, Boston.

Brown, C. M. S., Fitzgerald, K. F., Williams, D. G. & Maini, R. N. (1991). Immunoglobulin heavy chain variable region genes expressed by B cells in rheumatoid synovium. *American College of Rheumatology 55th Annual Scientific Meeting,* 17–21 November, Boston.

Buluwela, L. & Rabbitts, T. H. (1988). A VH gene is located within 95 kb of the human immunoglobulin heavy chain constant region genes. *Eur. J. Immunol.* **18**, 1843–1845.

Buluwela, L., Albertson, D. G., Sherrington, P., Rabbits, P. H., Spurr, N. & Rabbitts, T. H. (1988). The use of chromosomal translocations to study human immunoglobulin gene organization: mapping D_H segments within 35 kb of the C_μ gene and identification of a new D_H locus. *EMBO J.* **7**, 2003–2010.

Buluwela, L., Forster, A., Boehm, T. & Rabbitts, T. H. (1989). A rapid procedure for colony screening using nylon filters. *Nucl. Acids Res.* **17**, 452.

Bye, J. M., Carter, C., Cui, Y., Gorick, B. D., Songsivilai, S., Winter, G., Hughes-Jones, N. C. & Marks, J. D. (1992). *J. Clin. Invest.* in the press.

Cairns, E., Kwong, P. C., Misener, V., Ip, P., Bell, D. A. & Siminovitch, K. A. (1989). Analysis of variable region genes encoding a human anti-DNA antibody of normal origin. Implications for the molecular basis of human autoimmune responses. *J. Immunol.* **143**, 685–691.

Carroll, W. L., Yu, M., Lin, M. P. & Korsmeyer, S. J. (1989). Absence of Ig V region gene somatic hypermutation in advanced Burkitt's lymphoma. *J. Immunol.* **143**, 692–698.

Carter, P., Bedouelle, H. & Winter, G. (1985). Improved oligonucleotide site-directed mutagenesis using M13 vectors. *Nucl. Acids Res.* **13**, 4431–4443.

Chen, P. P. (1990). Structural analyses of human developmentally regulated Vh3 genes. *Scand. J. Immunol.* **31**, 257–267.

Chen, P. P. & Yang, P. M. (1990). A segment of human Vh gene locus is duplicated. *Scand. J. Immunol.* **31**, 593–599.

Chen, P. P., Liu, M. F., Sinha, S. & Carson, D. A. (1988). A 16/6 idiotype-positive anti-DNA antibody is encoded by a conserved VH gene with no somatic mutation. *Arthritis Rheum.* **31**, 1429–1431.

Chen, P. P., Liu, M. F., Glass, C. A., Sinha, S., Kipps, T. J. & Carson, D. A. (1989). Characterization of two immunoglobulin VH genes that are homologous to human rheumatoid factors. *Arthritis Rheum.* **32**, 72–76.

Chen, P. P., Olsen, N. J., Yang, P.-M., Soto-Gil, R. W., Olee, T., Siminovitch, K. A. & Carson, D. A. (1990). From human autoantibodies to the fetal antibody repertoire to B cell malignancy: its a small world after all. *Int. Rev. Immunol.* **5**, 239–251.

Chothia, C. & Lesk, A. M. (1987). Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* **196**, 901–917.

Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith, G. S., Air, G., Sheriff, S., Padlan, E. A., Davies, D., Tulip, W. R., *et al.* (1989). Conformations of immunoglobulin hypervariable regions. *Nature (London),* **342**, 877–883.

Chothia, C., Lesk, A. M., Gherardi, E., Tomlinson, I. M., Walter, G., Marks, J. D., Llewelyn, M. B. & Winter, G. (1992). The structural repertoire of human V_H segments. *J. Mol. Biol.* **227**, 799–817.

Cleary, M. L., Meeker, T. C., Levy, S., Lee, E., Trela, M., Sklar, J. & Levy, R. (1986). Clustering of extensive somatic mutations in the variable region of an immunoglobulin heavy chain gene from a human B cell lymphoma. *Cell,* **44**, 97–106.

Croce, C. M., Shander, M., Martinis, J., Cicurel, L., D'Ancona, G. G., Dolby, T. W. & Koprowski, H. (1979). Chromosomal location of the genes for human immunoglobulin heavy chains. *Proc. Nat. Acad. Sci., U.S.A.* **76**, 3416–3419.

Davidson, A., Manheimer, L. A., Aranow, C., Peterson, R., Hannigan, N. & Diamond, B. (1990). Molecular characterization of a somatically mutated anti-DNA antibody bearing two systemic lupus erythematosus-related idiotypes. *J. Clin. Invest.* **85**, 1401–1409.

Deane, M. & Norton, J. D. (1990). Immunoglobulin heavy chain variable region family usage is independent of tumor cell phenotype in human B lineage leukemias. *Eur. J. Immunol.* **20**, 2209–2217.

Denny, C. T., Yoshikai, Y., Mak, T. W., Smith, S. D., Hollis, G. F. & Kirsch, I. R. (1986). A chromosome 14 inversion in a T-cell lymphoma is caused by site-specific recombination between immunoglobulin and T-cell receptor loci. *Nature (London),* **320**, 549–551.

Dersimonian, H., Schwartz, R. S., Barrett, K. J. & Stollar, B. D. (1987). Relationship of human variable region heavy chain germ-line genes to genes encoding anti-DNA autoantibodies. *J. Immunol.* **139**, 2496–2501.

Dersimonian, H., McAdam, K. P., Mackworth, Y. C. & Stollar, B. D. (1989). The recurrent expression of variable region segments in human IgM anti-DNA autoantibodies. *J. Immunol.* **142**, 4027–4033.

Dersimonian, H., Long, A., Rubinstein, D., Stollar, B. D. & Schwartz, R. S. (1990). V_H genes of human autoantibodies. *Int. Rev. Immunol.* **5**, 253–264.

Desai, R., Spatz, L., Matsuda, T., Ilyas, A. A., Berman, J. E., Alt, F. W., Kabat, E. A. & Latov, N. (1990). Molecular cloning of a human immunoglobulin heavy chain variable (VH) region with anti-myelin-associated glycoprotein activity. *J. Neuroimmunol.* **26**, 35–41.

Dower, W. J., Miller, J. F. & Ragsdale, C. W. (1988). High efficiency transformation of E. coli by high voltage electroporation. *Nucl. Acids Res.* **16**, 6127–6145.

Ermel, R. W., Kenny, T. P. & Robbins, D. L. (1991). Molecular genetic characterization of human monoclonal IgM rheumatoid factors derived from rheumatoid synovial cells. *American College of Rheumatology 55th Annual Scientific Meeting,* 17–21 November, Boston.

Ezaki, I., Kanda, H., Sakai, K., Fukui, N., Shingu, M., Nobunaga, M. & Watanabe, T. (1991). Restricted diversity of the variable region nucleotide sequences of the heavy and light chains of a human rheumatoid factor. *Arthritis Rheum.* **34**, 343–350.

Felgenhauer, M., Kohl, J. & Ruker, F. (1990). Nucleotide sequences of the cDNAs encoding the V-regions of H- and L-chains of a human monoclonal antibody specific to HIV-1-gp41. *Nucl. Acids Res.* **18**, 4927.

Friedlander, R. M., Nussenzweig, M. C. & Leder, P. (1990). Complete nucleotide sequence of the membrane form of the human IgM heavy chain. *Nucl. Acids Res.* **18**, 4278.

Friedman, D. F., Cho, E. A., Goldman, J., Carmack, C. E., Besa, E. C., Hardy, R. R. & Silberstein, L. E. (1991). The role of clonal selection in the pathogenesis of an autoreactive human B cell lymphoma. *J. Exp. Med.* **174**, 525–537.

Frohman, M. A. & Martin, G. R. (1990). Detection of homologous recombinants. In *PCR Protocols: A Guide to Methods and Applications* (Innis, M. A., Gelfand, D. H., Sninsky, J. J. & White, T. J., eds), pp. 228–236, Academic Press, San Diego.

Geng, L., Silverman, E., Isacovics, B., Zheng, L., Dosch, H.-M. & Siminovitch, K. (1991). Characterization and sequence of human monoclonal anti-Ro/La antibodies. *American College of Rheumatology 55th Annual Scientific Meeting*, 17–21 November, Boston.

Gillies, S. D., Dorai, H., Wesolowski, J., Majeau, G., Young, D., Boyd, J., Gardner, J. & James, K. (1989). Expression of human anti-tetanus toxoid antibody in transfected murine myeloma cells. *Bio/Technology*, **7**, 799–804.

Gronenborn, B. (1976). Overproduction of phage lambda repressor under control of the lac promotor of Escherichia coli. *Mol. Gen. Genet.* **148**, 243–250.

Guillaume, T., Rubinstein, D. B., Young, F. Tucker, L., Logtenberg, T., Schwartz, R. S. & Barrett, K. J. (1990). Individual VH genes detected with oligonucleotide probes from the complementarity-determining regions. *J. Immunol.* **145**, 1934–1945.

Hillson, J. L. & Perlmutter, R. M. (1990). Autoantibodies and the fetal antibody repertoire. *Int. Rev. Immunol.* **5**, 215–29.

Hoch, S. & Schwaber, J. (1987). Identification and sequence of the VH gene elements encoding a human anti-DNA antibody. *J. Immunol.* **139**, 1689–1693.

Hughes-Jones, N. C., Bye, J. M., Beale, D. & Coadwell, J. (1990). Nucleotide sequences and three-dimensional modelling of the VH and VL domains of two human monoclonal antibodies specific for the D antigen of the human Rh-blood-group system. *Biochem. J.* **268**, 135–140.

Humphries, C. G., Shen, A., Kuziel, W. A., Capra, J. D., Blattner, F. R. & Tucker, P. W. (1988). A new human immunoglobulin VH family preferentially rearranged in immature B-cell tumours. *Nature (London)*, **311**, 446–449.

Kabat, E. A. & Wu, T. T. (1971). Attempts to locate complementarity-determining residues in the variable positions of light and heavy chains. *Ann. N.Y. Acad. Sci.* **190**, 382–393.

Kabat, E. A., Wu, T. T., Perry, H. M., Gottesman, K. S. & Foeller, C. (1991). *Sequences of Proteins of Immunological Interest*. 5th Edit. U.S. Department of Health and Human Services, Public Health Service National Institutes of Health.

Karr, N. S., Sasso, E. H., Hillson, J. L., Oppliger, I. R. & Mannik, M. (1991). Interaction between staphylococcal protein A and human antibodies of the VHIII family. *American College of Rheumatology 55th Annual Scientific Meeting*. 17–21 November, Boston.

Kenten, J. H., Molgaard, H. V., Houghton, M., Derbyshire, R. B., Viney, J., Bell, L. O. & Gould, H. J. (1982). Cloning and sequence determination of the gene for the human immunoglobulin epsilon chain expressed in a myeloma cell line. *Proc. Nat. Acad. Sci., U.S.A.* **79**, 6661–6665.

Kipps, T. J., Tomhave, E., Pratt, L. F., Duffy, S., Chen, P. P. & Carson, D. A. (1989). Developmentally restricted immunoglobulin heavy chain variable region gene expressed at high frequency in chronic lymphocytic leukemia. *Proc. Nat. Acad. Sci., U.S.A.* **86**, 5913–5917.

Kirsch, I. R., Morton, C. C., Nakahara, K. & Leder, P. (1982). Human immunoglobulin heavy chain genes map to a region of translocations in malignant B lymphocytes. *Science*, **216**, 301–303.

Kishimoto, T., Okajima, H., Okumoto, T. & Taniguchi, M. (1989). Nucleotide sequences of the cDNAs encoding the V-regions of H- and L-chains of a human monoclonal antibody with broad reactivity to malignant tumor cells. *Nucl. Acids Res.* **17**, 4385.

Kodaira, M., Kinashi, T., Umemura, I., Matsuda, F., Noma, T., Ono, Y. & Honjo, T. (1986). Organization and evolution of variable region genes of the human immunoglobulin heavy chain. *J. Mol. Biol.* **190**, 529–541.

Kon, S., Levy, S. & Levy, R. (1987). Retention of an idiotypic determinant in a human B-cell lymphoma undergoing immunoglobulin variable-region mutation. *Proc. Nat. Acad. Sci., U.S.A.* **84**, 5053–5057.

Kudo, A., Ishihara, T., Nishimura, Y. & Watanabe, T. (1985). A cloned human immunoglobulin heavy chain gene with a novel direct-repeat sequence in 5′ flanking region. *Gene*, **33**, 181–189.

Kuppers, R., Gause, A. & Rajewsky, K. (1991). B cells of chronic lymphatic leukemia express V genes in unmutated form. *Leuk. Res.* **15**, 487–496.

Larrick, J. W., Danielsson, L., Brenner, C. A., Abrahamson, M., Fry, K. E. & Borrebaeck, C. A. (1989*a*). Rapid cloning of rearranged immunoglobulin genes from human hybridoma cells using mixed primers and the polymerase chain reaction. *Biochem. Biophys. Res. Commun.* **160**, 1250–1256.

Larrick, J. W., Danielsson, L., Brenner, C. A., Wallace, E. F., Abrahamson, M., Fry, K. E. & Borrebaeck, C. A. K. (1989*b*). Polymerase chain reaction using mixed primers: cloning of human monoclonal antibody variable region genes from single hybridoma cells. *Bio/Technology*, **7**, 934–938.

Lee, K. H., Matsuda, F., Kinashi, T., Kodaira, M. & Honjo, T. (1987). A novel family of variable region genes of the human immunoglobulin heavy chain. *J. Mol. Biol.* **195**, 761–768.

Logtenberg, T., Young, F. M., Van, E. J., Gmelig, M. F. & Alt, F. W. (1989). Autoantibodies encoded by the most Jh-proximal human immunoglobulin heavy chain variable region gene. *J. Exp. Med.* **170**, 1347–1355.

Manheimer-Lory, A., Katz, J. B., Pillinger, M., Ghossein, C., Smith, A. & Diamond, B. (1991). Molecular characteristics of antibodies bearing an anti-DNA-associated idiotype. *J. Exp. Med.* **174**, 1639–1652.

Marks, J. D., Hoogenboom, H. R., Bonnert, T. P., McCafferty, J., Griffiths, A. D. & Winter, G. (1991*a*). By-passing immunization: human antibodies from V-gene libraries displayed on phage. *J. Mol. Biol.* **222**, 581–597.

Marks, J. D., Tristrem, M., Karpas, A. & Winter, G. (1991*b*). Oligonucleotide primers for polymerase chain reaction amplification of human immunoglobulin variable genes and design of family-specific oligonucleotide probes. *Eur. J. Immunol.* **21**, 985–991.

Maruyama, I. N. (1990). Estimation of errors in the

polymerase chain reaction. *Techniques-J. Meth. Cell. Mol. Biol.* **2**, 302–303.

Matsuda, F., Lee, K. H., Nakai, S., Sato, T., Kodaira, M., Zong, S. Q., Ohno, H., Fukuhara, S. & Honjo, T. (1988). Dispersed localization of D segments in the human immunoglobulin heavy-chain locus. *EMBO J.* **7**, 1047–1051.

Matsuda, F., Shin, E. K., Hirabayashi, Y., Nagaoka, H., Yoshida, M. C., Zong, S. Q. & Honjo, T. (1990). Organization of variable region segments of the human immunoglobulin heavy chain: duplication of the D_5 cluster within the locus and interchromosomal translocation of variable region segments. *EMBO J.* **9**, 2501–2506.

Matthyssens, G. & Rabbitts, T. H. (1980). Structure and multiplicity of genes for the human immnoglobulin heavy chain variable region. *Proc. Nat. Acad. Sci., U.S.A.* **77**, 6561–6565.

Meeker, T. C., Grimaldi, J. C., O'Rourke, R., Loeb, J., Juliusson, G. & Einhorn, S. (1988). Lack of detectable somatic hypermutation in the V region of the Ig H chain gene of a human chronic B lymphocytic leukemia. *J. Immunol.* **141**, 3994–3998.

Mensink, E. J., Schuurman, R. K., Schot, J. D., Thompson, A. & Alt, F. W. (1986). Immunoglobulin heavy chain gene rearrangements in X-linked agammaglobulinemia. *Eur. J. Immunol.* **16**, 963–967.

Mierau, R., Gause, A., Kuppers, R., Michels, M., Mageed, R. A., Jefferis, R. & Genth, E. (1992). A human monoclonal IgA rheumatoid factor using the VkIV light chain gene. *Rheum. Int.* **12**, 23–31.

Miller, J. H. (1972). *Experiments in Molecular Genetics*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Mills, D. R. & Kramer, F. R. (1979). Structure-independent nucleotide sequence analysis. *Proc. Nat. Acad. Sci., U.S.A.* **76**, 2232–2235.

Mortari, F., Ochs, H. D., Wedgwood, R. J. & Schroeder, H. J. (1991). Immunoglobulin variable heavy chain cDNA sequence from a patient with X-linked agammaglobulinemia. *Nucl. Acids Res.* **19**, 673.

Nakatani, T., Nomura, N., Horigome, K., Ohtsuka, H. & Noguchi, H. (1989). Functional expression of human monoclonal antibody genes directed against pseudomonal exotoxin A in mouse myeloma cells. *Bio/Technology*, **7**, 805–810.

Newkirk, M. M., Gram, H., Heinrich, G. F., Ostberg, L., Capra, J. D. & Wasserman, R. L. (1988). Complete protein sequences of the variable regions of the cloned heavy and light chains of a human anticytomegalovirus antibody reveal a striking similarity to human monoclonal rheumatoid factors of the Wa idiotypic family. *J. Clin. Invest.* **81**, 1511–1518.

Nickerson, K. G., Berman, J., Glickman, E., Chess, L. & Alt, F. W. (1989). Early human IgH gene assembly in Epstein–Barr virus-transformed fetal B cell lines. Preferential utilization of the most JH-proximal D segment (DQ52) and two unusual VH-related rearrangements. *J. Exp. Med.* **169**, 1391–1403.

Noma, Y., Yaoita, Y., Matsunami, N., Rosen, A., Klein, G. & Honjo, T. (1984). Immunoglobulin gene organization of ultraviolet-illuminated human lymphoblastoid cell lines producing both IgM and IgG. *Mol. Biol. Med.* **2**, 337–350.

Olee, T., Yang, P. M., Siminovitch, K. A., Olsen, N. J., Hillson, J., Wu, J., Kozin, F., Carson, D. A. & Chen, P. P. (1991). Molecular basis of an autoantibody-associated restriction fragment length polymorphism that confers susceptibility to autoimmune diseases. *J. Clin. Invest.* **88**, 193–203.

Pascual, V. & Capra, J. D. (1991). Human immunoglobulin heavy chain variable region genes: organization, polymorphism and expression. *Advan. Immunol.*, vol. 49, pp. 1–147, Academic Press, San Diego.

Pascual, V., Randen, I., Thompson, K., Sioud, M., Forre, O., Natvig, J. & Capra, J. D. (1990). The complete nucleotide sequences of the heavy chain variable regions of six monospecific rheumatoid factors derived from Epstein–Barr virus-transformed B cells isolated from the synovial tissue of patients with rheumatoid arthritis. Further evidence that some autoantibodies are unmutated copies of germ line genes. *J. Clin. Invest.* **86**, 1320–1328.

Pascual, V., Victor, K., Lelsz, D., Spellerberg, M. B., Hamblin, T. J., Thompson, K. M., Randen, I., Natvig, J., Capra, J. D. & Stevenson, F. K. (1991). Nucleotide sequence analysis of the V regions of two IgM cold agglutinins. Evidence that the VH4-21 gene segment is responsible for the major cross-reactive idiotype. *J. Immunol.* **146**, 4385–4391.

Perry, D. J. & Carrell, R. W. (1989). CpG dinucleotides are "hotspots" for mutation in the antithrombin III gene. Twelve variants identified using the polymerase chain reaction. *Mol. Biol. Med.* **6**, 239–243.

Rechavi, G., Bienz, B., Ram, D., Ben, N. Y., Cohen, J. B., Zakut, R. & Givol, D. (1982). Organization and evolution of immunoglobulin VH gene subgroups. *Proc. Nat. Acad. Sci., U.S.A.* **79**, 4405–4409.

Rechavi, G., Ram, D., Glazer, L., Zakut, R. & Givol, D. (1983). Evolutionary aspects of immunoglobulin heavy chain variable region (VH) gene subgroups. *Proc. Nat. Acad. Sci., U.S.A.* **80**, 855–859.

Reynaud, C.-A., Dahan, A., Anquez, V. & Weill, J.-C. (1989). Somatic hyperconversion diversifies the single V_H gene of the chicken with a high incidence in the D region. *Cell*, **59**, 171–183.

Rioux, J. D., Larose, Y., Brodeur, B., Rauch, J. E. & Newkirk, M. M. (1991). Structural relationships between rheumatoid factors and anti-viral antibodies. *American College of Rheumatology 55th Annual Scientic Meeting*, 17–21 November, Boston.

Robbins, D. L., Kenny, T. P., Coloma, M. J., Gavilondo, C. J., Soto, G. R., Chen, P. P. & Larrick, J. W. (1990). Serologic and molecular characterization of a human monoclonal rhematoid factor derived from rheumatoid synovial cells. *Arthritis Rheum.* **33**, 1188–1195.

Roudier, J., Silverman, G. J., Chen, P. P., Carson, D. A. & Kipps, T. J. (1990). Intraclonal diversity in the VH genes expressed by $CD5^-$ chronic lymphocytic leukemia-producing pathologic IgM rheumatoid factor. *J. Immunol.* **144**, 1526–1530.

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B. & Erlich, H. A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, **239**, 487–491.

Sam, M., Walter, M. A. & Cox, D. W. (1988). RsaI polymorphism of a human immunoglobulin VH5 subclass locus. *Nucl. Acids Res.* **16**, 8748.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Nat. Acad. Sci., U.S.A.* **74**, 5463–5467.

Sanz, I., Casali, P., Thomas, J. W., Notkins, A. L. & Capra, J. D. (1989*a*). Nucleotide sequences of eight human natural autoantibody Vh regions reveals apparent restricted use of VH families. *J. Immunol.* **142**, 4054–4061.

Sanz, I., Dang, H., Takei, M., Talal, N. & Capra, J. D.

(1989*b*). V_H sequence of a human anti-Sm autoantibody. Evidence that autoantibodies can be unmutated copies of germline genes. *J. Immunol.* **142**, 883–887.

Sanz, I., Kelly, P., Williams, C., Scholl, S., Tucker, P. & Capra, J. D. (1989*c*). The smaller human VH gene families display remarkably little polymorphism. *EMBO J.* **8**, 3741–3748.

Sasso, E. H., van Dijk, K. W. & Milner, E. C. (1990). Prevalence and polymorphism of human VH3 genes. *J. Immunol.* **145**, 2751–2757.

Sasso, E. H., Silverman, G. J. & Mannik, M. (1991). Human IgA and IgG F(ab')2 that bind to staphylococcal protein A belong to the VHIII subgroup. *J. Immunol.* **147**, 1877–1883.

Schroeder, H. J. & Wang, J. Y. (1990). Preferential utilization of conserved immunoglobulin heavy chain variable gene segments during human fetal life. *Proc. Nat. Acad. Sci., U.S.A.* **87**, 6146–6150.

Schroeder, H. J., Hillson, J. L. & Perlmutter, R. M. (1987). Early restriction of the human antibody repertoire. *Science,* **238**, 791–793.

Schroeder, H. J., Hillson, J. L. & Perlmutter, R. M. (1990). Structure and evolution of mammalian VH families. *Int. Immunol.* **2**, 41–50.

Schutte, M. E., Ebeling, S. B., Akkermans, K. E., Gmelig, M. F. & Logtenberg, T. (1991). Antibody specificity and immunoglobulin VH gene utilization of human monoclonal CD5+ cell lines. *Eur. J. Immunol.* **21**, 1115–1121.

Shen, A., Humphries, C., Tucker, P. & Blattner, F. (1987). Human heavy-chain variable region gene family nonrandomly rearranged in familial chonic lymphocytic leukemia. *Proc. Nat. Acad. Sci., U.S.A.* **84**, 8563–8567.

Shin, E. K., Matsuda, F., Nagaoka, H., Fukita, Y., Imai, T., Yokoyama, K., Soeda, E. & Honjo, T. (1991). Physical map of the 3′ region of the human immunoglobulin heavy chain locus: clustering of autoantibody-related variable segments in one haplotype. *EMBO J.* **10**, 3641–3645.

Shuldiner, A. R., Nirula, A. & Roth, J. (1989). Hybrid DNA artifact from PCR of closely related target sequences. *Nucl. Acids Res.* **17**, 4409.

Silberstein, L. E., Litwin, S. & Carmack, C. E. (1989). Relationship of variable region genes expressed by a human B cell lymphoma secreting pathologic anti-Pr2 erythrocyte autoantibodies. *J. Exp. Med.* **169**, 1631–1643.

Silberstein, L. E., Jefferies, L. C., Goldman, J., Friedman, D., Nowell, P. C., Roelcke, D., Pruzanski, W., Roudier, J. & Silverman, G. J. (1991). Variable region gene analysis of pathologic human autoantibodies to the related i and I red blood cell antigens. *Blood,* **78**, 2372–2386.

Siminovitch, K. A. & Chen, P. P. (1990). The biologic significance of human natural autoimmune responses; relationship to the germline, early immune and malignant B cell variable gene repertoire. *Int. Rev. Immunol.* **5**, 265–277.

Souroujon, M. C., Rubinstein, D. B., Schwartz, R. S. & Barrett, K. J. (1989). Polymorphisms in human H chain V region genes from the VHIII gene family. *J. Immunol.* **143**, 706–711.

Spatz, L. A., Wong, K. K., Williams, M., Desai, R., Golier, J., Berman, J. E., Alt, F. W. & Latov, N. (1990). Cloing and sequence analysis of the VH and VL regions of an anti-myelin/DNA antibody from a patient with peripheral neuropathy and chronic lymphocytic leukemia. *J. Immunol.* **144**, 2821–2828.

Takahashi, N., Noma, T. & Honjo, T. (1984). Rearranged immunoglobulin heavy chain variable region (VH) pseudogene that deletes the second complementarity-determining region. *Proc. Nat. Acad. Sci., U.S.A.* **81**, 5194–5198.

Timmers, E., Kenter, M., Thompson, A., Kraakman, M. E. M., Berman, J. E., Alt, F. W. & Schuurman, R. K. B. (1991). Diversity of immunoglobulin heavy chain gene segment rearrangement in B lymphoblastoid cell lines from X-linked agammaglobulinemia patients. *Eur. J. Immunol.* **21**, 2355–2363.

Tonegawa, S. (1983). Somatic generation of antibody diversity. *Nature (London),* **302**, 575–581.

Turnbull, I. F., Bernard, O., Sriprakash, K. S. & Mathews, J. D. (1987). Human immunoglobulin variable region genes: a new VH sequence used to detect polymorphism. *Immunogenetics,* **25**, 184–192.

van der Heijden, R. W. J., Bunschoten, H., Pascual, V., Uytdehaag, F. G. C. M., Osterhaus, A. D. M. E. & Capra, J. D. (1990). Nucleotide sequence of a human monoclonal anti-idiotypic antibody specific for a rabies virus-neutralizing monoclonal idiotypic antibody reveals extensive somatic variability suggestive of an antigen driven immune response. *J. Immunol.* **144**, 2835–2839.

van Dijk, K. W., Sasso, E. H. & Milner, E. C. (1991). Polymorphism of the human Ig VH4 gene family. *J. Immunol.* **146**, 3646–3651.

van Es, J. H., Gmelig Meyling, F. H. J., van de Akker, W. R. M., Aanstoot, H., Derksen, R. H. W. M. & Logtenberg, T. (1991). Somatic mutations in the variable regions of a human IgG anti-double-stranded DNA autoantibody suggest a role for antigen in the induction of systemic lupus erythematosus. *J. Exp. Med.* **173**, 461–470.

Walter, M. A., Surti, U., Hofker, M. H. & Cox, D. W. (1990). The physical organization of the human immunoglobulin heavy chain gene complex. *EMBO J.* **9**, 3303–3313.

White, M. B., Word, C. J., Humphries, C. G., Blattner, F. R. & Tucker, P. W. (1990). Immunoglobulin D switching can occur through homologous recombination in human B cells. *Mol. Cell. Biol.* **10**, 3690–3699.

Winter, G. & Milstein, C. (1991). Man-made antibodies. *Nature (London),* **349**, 293–299.

Wood, W. I., Gitschier, J., Lasky, L. A. & Lawn, R. M. (1985). Base composition-independent hybridization in tetramethylammonium chloride: a method for oligonucleotide screening of highly complex gene libraries. *Proc. Nat. Acad. Sci., U.S.A.* **82**, 1585–1588.

Wysocki, L. J. & Gefter, M. L. (1989). Gene conversion and the generation of antibody diversity. *Annu. Rev. Biochem.* **58**, 509–531.

Yasui, H., Akahori, Y., Hirano, M., Yamada, K. & Kurosawa, Y. (1989). Class switch from μ to δ is mediated by homologous recombination between $\sigma\mu$ and $\Sigma\mu$ sequences in human immunoglobulin gene loci. *Eur. J. Immunol.* **19**, 1399–1403.

Edited by J. Karn

Note added in proof. Since submission of this paper, we have amplified and cloned six additional V_H segments from DP (DP-75 to DP-80). EMBL Data Library accession numbers for DP-1 to DP-80; Z12303-37, Z12602-3, Z12338-74 and Z14071-6.